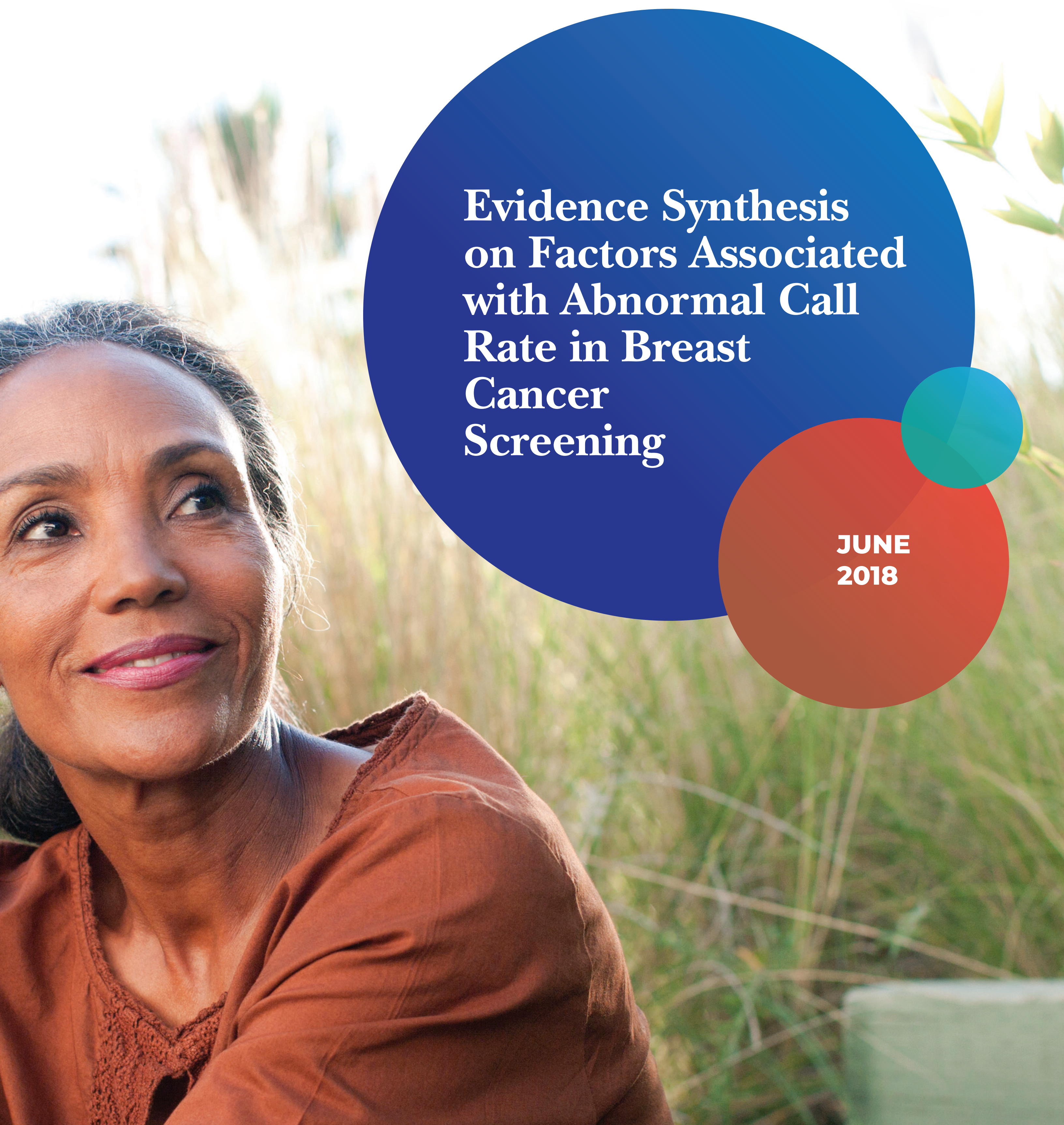
CANADIAN PARTNERSHIP AGAINST CANCER
PARTENARIAT CANADIEN CONTRE LE CANCER



Evidence Synthesis on Factors Associated with Abnormal Call Rate in Breast Cancer Screening

JUNE 2018

This work was undertaken to directly inform the development of the ***Pan-Canadian Framework for Action to Address Abnormal Call Rates in Breast Cancer Screening*** led by Partnership.

Production of this report has been made possible through a financial contribution from Health Canada, through the Canadian Partnership Against Cancer.

If you wish to use the information therein please ensure appropriate attribution by referencing it: *Research Sciences International, Canadian Partnership Against Cancer. Evidence Synthesis on Factors Associated with Abnormal Call Rate in Breast Cancer Screening. Prepared for: Canadian Partnership Against Cancer*. 2018 Jun 22.

Canadian Partnership Against Cancer
145 King Street, Suite 900, Toronto ON M5H 1J8
Phone: 416-915-9222, Fax: 416-915-9224

For more information on this publication, please email:
screening@partnershipagainstcancer.ca

Contents

Executive Summary

The Canadian Partnership Against Cancer (the Partnership) is interested in examining literature on factors potentially associated with the abnormal call rate in Canada. The findings are intended to be used by the Partnership to inform discussions with stakeholders on the factors affecting the abnormal call rate in Canada and how they can be addressed in practice.

OBJECTIVES: The specific objective of this report is to determine which factors are associated with the increasing trends seen in abnormal call rates. More specifically, the primary research questions to be addressed are given below.

- Is the transition from film-screen mammography to digital mammography (computed radiography and digital radiography) responsible for the increasing abnormal call rate?
- Are differences in quality assurances practices in breast cancer screening (minimum number of annual reads by a radiologist, approaches to double-reading, etc.) associated with the increasing abnormal call rate?
- Are differences in radiologist training or the use of computer-aided detection systems associated with the increasing abnormal call rate?
- Are radiologist characteristics (e.g., gender, experience, values, and concerns such as litigation) associated with the increasing abnormal call rate?

METHODS: To address these objectives, a synthesis of the relevant literature was conducted using four bibliographic databases: Medline, Embase, Scopus, and Cochrane Database of Systematic Reviews. The primary search results were supplemented with articles identified from reference lists. Search results were screened by title and abstract using pre-defined eligibility criteria developed in consultation with the Partnership. Recent relevant review articles of reasonable quality were used as a starting point for data synthesis. The search date reported in the systematic review selected for each factor was used as the starting date for inclusion of more recent original publications not covered by this systematic review. In preparation for the tabular summaries of key findings, a data abstraction form was developed and revised according to suggestions from the Partnership. Information extracted included data on: study and participant characteristics (program/study name, study period, target age, screening frequency, sample size, and participant age), potential influencing factors of recall rate (factor under study, and others related to mammography technology, quality assurance practices, or radiologist characteristics), quantitative results (recall rate, false positive rate, cancer detection rate, and positive predictive value), author reported limitations and conclusions, as well as any additional comments. A search of grey literature published by relevant Canadian and international associations was performed to supplement relevant reviews and key recent contributions captured by the current search strategy. The grey literature search focused on quality assurance practices adopted by breast cancer screening programs in different jurisdictions to enable the comparison of these practices in jurisdictions with lower and higher abnormal call rates.

RESULTS: Recall rates in Canada and the USA are higher than those in European countries or in Australia. At the same time, cancer detection rates in the USA and Canada are comparable to those in Europe and

lower than those in Australia. There are differences in quality assurance practices between breast cancer screening programs with high recall rates (Canada and USA) and programs with lower recall rates (Europe and Australia). In Australia and Europe, double reading of mammograms is an accepted practice while in the USA and Canada, double reading is not a standard practice. The minimum reading volume by a radiologist required by the European guidelines is 5,000 mammograms per year. The Australian guidelines require reading of at least 2,000 mammograms. In the USA and Canada, a minimum of 480 reads per year is accepted. In Australia, quarterly reporting of individual screen readers' performance is practiced: a quarterly Quality Assurance report includes the reader's recall to assessment rate. In the UK and Australia, screen readers can interpret a standard set of mammograms through a web-based software and receive immediate feedback on their performance. In Canada, accreditation is voluntary whereas in Australia it is mandatory. This comparison in different jurisdictions does not allow conclusions to be made regarding potential associations between adopted quality assurance practices and performance indicators. To further assess associations between potential influential factors (mammography technology, quality assurance practices and radiologist's characteristics) and performance indicators, review articles and original research were assessed.

Results from narrative and systematic reviews, as well as publications describing original research, are summarized in Table 1 (page 7).

SUMMARY OF MAIN FINDINGS: Based on the present review, the current evidence on factors potentially affecting breast cancer screening recall rates may be summarized as follows.

Factors that may decrease recall rates without compromising cancer detection

- Implementation of digital breast tomosynthesis (DBT) in screening practice
- Targeted double reading of only potential recalls
- Consensus or arbitration as compared to unilateral recall at double reading
- Comparison with two or more prior mammograms
- Batch reading of mammograms
- Fellowship training in breast imaging

Factors that may merit further consideration in designing breast cancer screening programs

- Synthesized mammography. This factor was not analyzed in depth within the framework of this project. However, initial screen of literature suggests that, used as an adjunct to DBT, this technology may preserve the performance benefits provided by DBT and at the same time reduce the dose of radiation.
- Interventions that include performance feedback and educational components are potentially effective in decreasing recall rates while maintaining cancer detection rates. Factors that determine their effectiveness need to be identified.
- Reading volume: Although overall evidence is inconsistent, a Canadian study of good quality demonstrates gains in interpretive accuracy with increasing reading volume; the gain is greater in the range of reading volumes up to about 3000 mammograms per year.

- Mammographic compression: evidence from a single study shows that false positive rates may be lower and cancer detection rates are significantly higher at moderate compression pressure compared to low or high pressure.

Non-modifiable factors that may influence recall rates

- Recall rates may decrease with increasing years of experience interpreting mammograms
- Female radiologists tend to have higher recall rates than male radiologists.

Factors with inconsistent or insufficient evidence on their effect on recall rates

- Introduction of digital mammography
- Computer assisted detection systems (CAD). There are different ways in which CAD is used (e.g., as a second reader, as an arbitrator of discordant opinions), and it can be used as an adjunct to different technologies. The effect of CAD on performance may differ depending on the way it is used as well as on the experience of screen reader who is using it. Little research is available to address these aspects of CAD use. It should also be noted that manufacturers of CAD systems work on improvements of CAD algorithms to increase specificity by reducing false prompts.

Table 1. Summary of findings on factors affecting abnormal call rate in breast cancer screening.

Factor	Sub-Factor	Review Articles	Original Research	Overall Conclusion	Comments
Mammography Technology	**Screen-film vs. digital mammography**	**Recall (false positive) rate** • No significant difference; significant heterogeneity reported among studies **Cancer detection rate** • Digital is better (or comparable) to screen film	**Recall rate** • Inconsistent results **Cancer detection rate** Inconsistent results	**Recall rate** • No clear evidence **Cancer detection rate** • No clear evidence	
	Computer-Aided Detection Systems	**Recall (false positive) rate** • Single reading with CAD vs. single reading: may increase the recall rate • Single reading with CAD vs. double reading: insufficient evidence **Cancer detection rate** • Single reading with CAD vs. single reading: no clear evidence of an effect • Single reading with CAD vs. double reading: insufficient evidence	**Recall rate** • Single reading with CAD vs. single reading: Increase (1 study) or unchanged (1 study) • Single reading with CAD vs. double reading: Inconsistent results • SFM with CAD vs. SFM: may increase recall rate (based on unadjusted results from 1 study) • DM with CAD vs. DM: No significant difference (based on 1 study) • CAD vs. third reader as arbitrator: may increase recall rate (based on 1 study) **Cancer detection rate** • Single reading with CAD vs. single reading: Increase (1 study) or unchanged (1 study)	**Recall rate** • Single reading with CAD vs. single reading: May increase the recall rate • Single reading with CAD vs. double reading: no clear evidence • SFM with CAD vs. SFM: Insufficient evidence • DM with CAD vs. DM: Insufficient evidence • CAD vs. third reader as arbitrator: Insufficient evidence **Cancer detection rate** • Single reading with CAD vs. single	•

Factor	Sub-Factor	Review Articles	Original Research	Overall Conclusion	Comments
			• Single reading with CAD vs. double reading: inconsistent results • SFM with CAD vs. SFM: No significant difference (based on 1 study) • DM with CAD vs. DM: No significant difference for cancer detection and invasive cancer detection rate. May increase DCIS detection rate. (based on 1 study) • CAD vs. third reader as arbitrator: Similar between groups being compared. CAD recalled one more case of cancer. (based on 1 study)	reading: no clear evidence • Single reading with CAD vs. double reading: no clear evidence • SFM with CAD vs. SFM: Insufficient evidence • DM with CAD vs. DM: Insufficient evidence • CAD vs. third reader as arbitrator: Insufficient evidence	
	Tomosynthesis	**Recall (false positive) rate** • May decrease **Cancer detection rate** • May increase	**Recall (false positive) rate** • Decreased in most studies **Cancer detection rate** • Increased in most studies	**Recall (false positive) rate** • May decrease **Cancer detection rate** • May increase	• May increase recalls for specific breast abnormalities • Benefits of tomosythesis may vary by patient's characteristics
	Synthesized mammography [in conjunction with tomosynthesis]	**Recall (false positive) rate** • May decrease **Cancer detection rate** • May increase or remain unchanged	**Recall (false positive) rate** Tomosynthesis + synthesized mammography vs. digital mammography • Inconsistent results Tomosynthesis + synthesized mammography	• May be beneficial as it appears to preserve the performance benefits provided by tomosynthesis and reduced the dose of radiation	• Only abstracts of relevant publications were reviewed. In depth analysis of this factor is required.

Factor	Sub-Factor	Review Articles	Original Research	Overall Conclusion	Comments
			vs. tomosynthesis+digital mammography • Decreased in most studies **Cancer detection rate** • Unchanged in most studies		
Quality Assurance Practices	**Reading volume**	**Recall (false positive) rate** • Higher reading volume may decrease recall rates (narrative reviews only) **Cancer detection rate** • There may be a threshold over which there is a decline in CDR (narrative reviews only)	**Recall (false positive) rate** • Inconsistent results **Cancer detection rate** • Inconsistent results	**Recall (false positive) rate** • No clear evidence **Cancer detection rate** • No clear evidence	A Canadian study of good quality shows a decrease in FP with increasing reading volume; greater gains in overall accuracy with increasing volume up to ≈3000 mammograms/year
	Double reading	**Recall (false positive) rate** • Arbitration/consensus may decrease • Unilateral practice may increase recall rates • Blinded double reading may increase false positives as compared to non-blinded double reading. **Cancer detection rate** • Double reading vs. single reading may increase CDR (unclear if improvement is dependent on method of resolution)	**Recall (false positive) rate** • Increased at double reading with consensus, arbitration or unilateral recalls vs. single reading. • Decreased at targeted double reading of only potential recalls vs. single reading • Increased at blinded vs. non-blinded double reading • Decreased at double reading with consensus/arbitration vs. double reading with unilateral recall	**Recall (false positive) rate** • No clear evidence for double reading with consensus or arbitration vs. single reading • May decrease at targeted double reading of only potential recalls vs. single reading • May increase at blinded vs. non-blinded double reading	Practices of double reading, recall strategies and readers' characteristics differ among mammography programs; generalizations of study results may be problematic. Benefits from double reading may be larger for less

Factor	Sub-Factor	Review Articles	Original Research	Overall Conclusion	Comments
		• Blinded reading may increase cancer detection. • Inconsistent finding regarding double reading with arbitration vs. unilateral recall	• Unchanged with changing the order in which the two readers examine <u>a batch</u> of mammograms **Cancer detection rate** • Increased or unchanged at double vs. single reading • Increased at blinded vs. non-blinded double reading • Not appreciably changed at double reading with consensus/arbitration vs. double reading with unilateral recall • Unchanged with changing the order in which thee two readers examine <u>a batch</u> of mammograms	• May decrease at double reading with consensus/arbitration vs. double reading with unilateral recall • May remain unchanged with changing the order in which the two readers examine <u>a batch</u> of mammograms **Cancer detection rate** • May increase or remain unchanged at double vs. single reading • May increase at blinded vs. non-blinded double reading • May not be significantly different at double reading with consensus/arbitration vs. double reading with unilateral recall • May remain unchanged with changing the order in which the two readers examine <u>a batch</u> of mammograms	experienced screen readers.

Factor	Sub-Factor	Review Articles	Original Research	Overall Conclusion	Comments
	Audit and performance feedback	• One review article discusses strengths and limitations of different methods to provide feedback but does not address the effect on recall or cancer detection rates.	• Two studies of the effectiveness of interventions that included performance feedback and educational components; one demonstrated a positive effect of a long-term intervention; the other demonstrated lack of effect of a short-term intervention on interpretive performance.	• Insufficient evidence	• Based on the available limited evidence, it is not possible to identify factors that determine the effectiveness of an intervention.
	Comparisons with prior mammograms	• No reviews identified	**Recall (false positive) rate** • Decreased; comparisons of the current mammogram with two ore more prior mammograms more effective than comparisons with a single prior mammogram **Cancer detection rate** • No evidence of negative impact of this practice on cancer detection	**Recall (false positive) rate** • May decrease, especially with two or more prior mammograms **Cancer detection rate** • May not be negatively affected	The proportion of prevalent screens is higher among women with no prior mammograms, which should be accounted for if women with no prior mammograms are used as a reference group.
	Number of mammographic views	**False positives**: • Two-view mammography decreases false positives vs. single-view (narrative review)	• Irrelevant factor; original research was not reviewed	• Irrelevant factor	Because screening mammography in Canada has been conducted using two views since the 1980's, it is unlikely that this practice contributed to the

Factor	Sub-Factor	Review Articles	Original Research	Overall Conclusion	Comments
					higher false positive rates in North America
	Mammographic compression	• No reviews identified	**Recall (false positive) rate** • False positives non-significantly decreased at moderate compression pressure vs. low or high compression pressure; no trend in recall rates (evidence from a single study) **Cancer detection rate** • Significantly increased at moderate compression pressure vs. low or high compression pressure (evidence from a single study)	**Recall (false positive) rate** • Insufficient evidence **Cancer detection rate** • Insufficient evidence	
	Batch reading of mammograms	• No reviews identified	**Recall (false positive) rate** • Decreased compared to immediate (online) non-batch reading in the presence of the patient and compared to offline non-batch reading after the patient left the premises **Cancer detection rate** • Unaffected by this practice	**Recall (false positive) rate** • May decrease **Cancer detection rate** • May not be negatively affected	One study specified that batch reading sessions were undertaken in uninterrupted distraction free environment.

Factor	Sub-Factor	Review Articles	Original Research	Overall Conclusion	Comments
Radiologist Characteristics	**Training, education, experience**	**False-positive rates**: • Fellowship training in breast imaging reduces false positive recall rates (narrative review)	**Recall (false positive) rate** • Decreased with increasing years of practice as demonstrated in most (but not all) studies • Fellowship-trained radiologists may not have a learning curve and achieve the performance goal within the first year (evidence from a single study) • No significant trend with time spent/hours per week working in breast imaging (evidence from a single study) • Inconsistent evidence regarding affiliation with an academic medical center • Radiologists who previously worked with tomosynthesis have higher recall rates working with digital mammography (evidence from a single study) **Cancer detection rate** • Radiologists who previously worked with tomosynthesis have higher cancer detection working with digital	**Recall (false positive) rate** • May decrease with increasing years of practice • May decrease in fellowship trained radiologists at earlier stages in their careers **Cancer detection rate** • Insufficient evidence	Comparisons between the groups defined by length of service can be confounded by changes in medical education and practices over time and by differences between radiologists who had decided to stay in mammography for many years and those who had recently entered the field

Factor	Sub-Factor	Review Articles	Original Research	Overall Conclusion	Comments
			mammography (evidence from a single study) • Insufficient evidence for other training and experience related factors (not reported in most studies)		
	Demographics	• No reviews identified	**Recall (false positive) rate** • Decrease with age observed in several studies is most likely due to increasing years of experience • Increased in female radiologists as demonstrated in two studies; analyses adjusted for other radiologists' characteristics such as experience **Cancer detection rate** • Insufficient evidence (not reported in most studies)	**Recall (false positive) rate** • May increased in female radiologists **Cancer detection rate** • Insufficient evidence	
	Litigation concerns	**Recall rates:** • May increase recall rates (narrative review)	**Recall (false positive) rate** • No evidence of an effect **Cancer detection rate** • Insufficient evidence (not reported in most studies)	**Recall (false positive) rate** • May not be affected **Cancer detection rate** • Insufficient evidence	Although radiologists reported being extremely concerned about medical malpractice and believed this concern affected their recall rates, variables

Factor	Sub-Factor	Review Articles	Original Research	Overall Conclusion	Comments
					characterizing medical malpractice experience and concerns were not associated with recall or false-positive rates

Definitions:

Inconsistent results: different outcomes reported for the factor of study.

Insufficient evidence: little or no research available.

No clear evidence: different conclusions reported between the reviews and original research article; or neither reviews nor original research provide consistent results.

Background

The Canadian Breast Cancer Screening Network (CBCSN) is responsible for supporting continuous improvement of breast cancer screening programs across Canada through collection, analysis, and interpretation of data on national quality indicators. Recent reports show that the national abnormal call rate for screening mammography rose by 23% between 2007 and 2012, while the invasive cancer detection rate was stable. This suggests that there is a larger proportion of women being called back for follow-up diagnostic testing who do not have cancer. This may cause potential harm to the individual, as well as requiring additional system resources. Besides these negative outcomes, the abnormal call rate has not met the national targets (<10% from initial screens and <5% for subsequent screens). These results merit further examination. The first phase of this project will examine potential factors that are related to the increasing abnormal call rate.

Risk Sciences International (RSI) was contracted to provide support to the Partnership through examining published literature on the factors associated with the abnormal call rate in Canada and other jurisdictions. The findings are intended to be used by the Partnership to inform discussions with stakeholders on the factors affecting the abnormal call rate in Canada and how they can be addressed. The intent is focused on knowledge mobilization or putting knowledge into practice amongst the key stakeholders. The primary audience for this work is the Partnership’s abnormal call rate project team, radiologists and program directors, with the secondary audience being the Canadian Breast Cancer Screening Network.

Approach and Methodology

Literature search strategy and selection of articles for review

Search in bibliographic databases

The search strategy was based on four concepts as outlined in Figure 1; specifically: (1) breast cancer, (2) screening, (3) abnormal call rate, and (4) potential influential factors (technology, quality assurance practices, and radiologist characteristics).

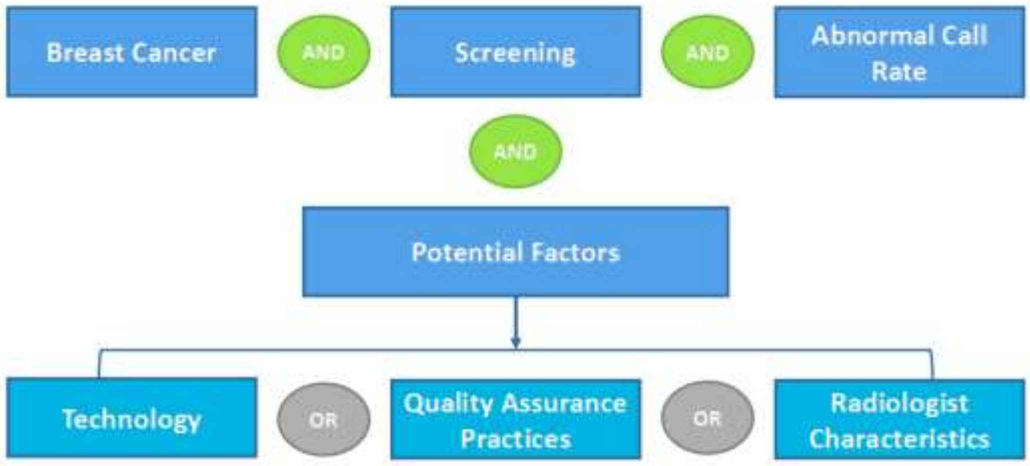


Figure 1. Concepts used in developing the literature search strategy for the present project.

Four electronic literature databases were consulted during the conduct of this work: Medline, Embase, Scopus, and the Cochrane Database of Systematic Reviews. Details on search terms used in each database are presented in Appendix 1. Since there is a significant (98%)[1] overlap between PubMed and Medline, and PubMed allows only limited control over search terms[2], a literature search in PubMed was not performed. As described in Figure 2, the initial search produced 991 results. These results were imported into an Endnote database, and 429 duplicate records were removed, leaving 562 citations in the database.

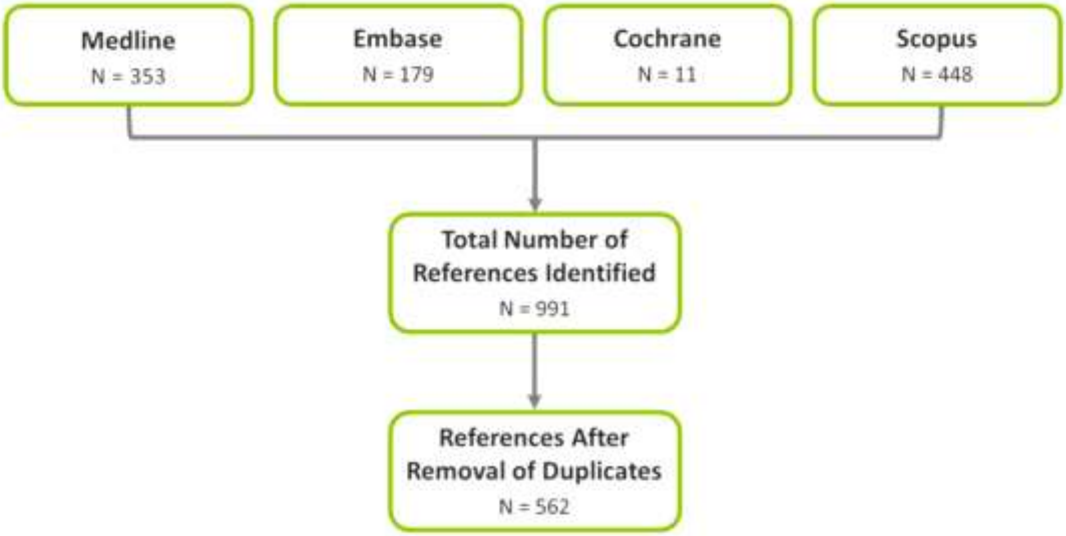

Medline
N = 353
Embase
N = 179
Cochrane
N = 11
Scopus
N = 448
Total Number of References Identified
N = 991
References After Removal of Duplicates
N = 562

Figure 2. Flow diagram illustrating the search results from the applied search strategy.

[1] See, for example: https://kemh.libguides.com/library/search_tips/faqs/difference_between_pubmed_medline_embase
[2] In particular, PubMed does not support adjacency searching: https://support.nlm.nih.gov/link/portal/28045/28054/Article/473/Does-PubMed-support-adjacency-searching

Eligibility criteria

The articles identified through the literature search were subjected to Level 1 (title and abstract) and Level 2 (full text) screening for relevance using eligibility criteria listed in Table 1.

Table 2. Eligibility criteria applied in screening potentially relevant literature.

Included	Excluded
• articles describing the effect of breast cancer screening technology on the rate of abnormal calls (including technology such as tomosynthesis, which is used for screening average-risk populations) • articles describing the effect of radiologist characteristics on the rate of abnormal calls • articles describing the effects of quality assurance practices on the rate of abnormal calls (for example: performance feedback, accreditation; and number of annual reads) • articles describing studies conducted in Canada, U.S., Europe, U.K., Australia, New Zealand • articles published 2003-present	• articles on diagnostic mammography [the focus of this project is on screening, rather than diagnostic mammography] • studies of breast imaging in selected high-risk populations (for example: women with BRCA1/BRCA2 mutations, breast cancer in first degree relatives, or high breast density) • studies of imaging techniques used as an adjunct to mammography (for example: MRI and ultrasound are excluded as these procedures are used primarily for diagnosis and for screening in selected high-risk populations) • articles describing studies conducted in countries other than Canada, U.S., Europe, U.K., Australia, New Zealand • articles published before 2003

Review articles

Results from identified review articles were used as a starting point for the evidence synthesis. Systematic reviews of studies on factors potentially influencing the abnormal call rates were scored for quality using the AMSTAR tool (**Shea et al. 2017**). For each factor, we selected the most relevant and comprehensive systematic review of reasonable quality (as described in Appendix 2). It should be noted that the AMSTAR tool is designed to assess not only the quality of implementation of a systematic review, but also the level of completeness of reporting on its methodology. We believe that the 'Critically Low-Quality" score assigned to most review articles identified for this report reflects low-quality reporting rather than implementation. The search date reported in the systematic review selected for each factor was used as the starting date for inclusion of more recent original publications not covered by the systematic review in order to bring the current literature on each factor up to date within the context of the evidence synthesis.

Data abstracted from review articles identified as relevant to this project are provided in Appendix 3.

Selection of peer reviewed publications reporting on original research results

Selection of original publications conducted as follows:

1) The inclusion dates for selecting original publications were modified for each factor to include more recent publications not covered by the systematic reviews. The modified inclusion dates

are listed in table 2. More details on how the inclusion dates were determined can be found in Appendix 2.

2) Recent original publications were screened using the eligibility criteria listed in Table 2.

Table 2. Identification of start dates for the selection of original publications using relevant systematic reviews.

Factor		Inclusion dates
Technology	Screen-film vs. Digital Mammography	2009-present
	Computer-Aided Detection Systems	(Single Reading + CAD vs. Single Reading) 2008-present (Single Reading + CAD vs. Double Reading) 2011-present
	Tomosynthesis	2014-present
	Synthesized Digital Mammography	2003-present
Quality Assurance Practices	Reading Volume	2003-present
	Double Reading	2003-present
	Audit/Performance Feedback	2003-present
	Comparison with Prior Mammograms	2003-present
	Number of Mammographic Views	2003-present
	Mammographic Compression	2003-present
	Other Quality Assurance Practices	2003-present
Radiologist Characteristics	Training, Education, and Experience	2003-present
	Age and gender	2003-present
	Litigation concern	2003-present

Articles describing interventions aimed at reducing recall rates while maintaining an acceptable cancer detection rate were included. Although the keyword “intervention” was not used in the literature searches, articles describing interventions, if identified through our original searches, were considered eligible for inclusion. Peer reviewed publications selected after title and abstract screening are listed in Appendix 4.

Grey literature

A search of grey literature published by relevant Canadian and international associations was performed to supplement relevant reviews and key recent contributions captured by the current search strategy. Our grey literature search focused on quality assurance practices adopted by breast cancer screening programs in different jurisdictions to enable the comparison of these practices in jurisdictions with lower and higher abnormal call rates. Literature describing interventions aimed at improving breast cancer screening performance indicators was also eligible for inclusion.

Data abstraction from publications on original research

In preparation for populating tabular summaries of key findings from original research, a data abstraction form was developed and revised according to suggestions from the Partnership.

Information extracted included data on: study and participant characteristics (program/study name, study period, target age, screening frequency, sample size, and participant age), potential influencing factors of recall rate (factor under study, and others related to mammography technology, quality assurance practices, or radiologist characteristics), quantitative results (recall rate, false positive rate, cancer detection rate, and positive predictive value), author reported limitations and conclusions, as well as any additional comments.

Title and abstract screening of literature search results produced a list of 90 publications (provided in Appendix 4). This list was subjected to further screening aimed at selecting the most informative studies and avoiding overrepresentation of studies conducted in the USA. Preference was given to publications based on “real world” data rather than on data obtained from test sets. Studies investigating the possible influence of synthesized mammography were not initially included in the scope of this project. For this reason, data from studies of synthesized mammography were not abstracted; a summary based on data from abstracts is provided. Tabular summaries of data from selected publications can be found in Appendix 5, along with reasons for exclusion of studies not retained for further analysis.

It has been reported that the risk of false-positive screening results is positively correlated with recall rates (Otten et al. 2005 as cited in Mohd Norsuddin et al. 2015). Thus, in the absence of data on recall rates, false-positive rates were abstracted. For some potential influencing factors, RSI could identify little or no information on either recall or false-positive rates. In this case, data on specificity[3] were abstracted. Likewise, if little or no information was identified on cancer detection rates and/or positive predictive value of recall, RSI abstracted data on sensitivity. Cancer detection rate is related to sensitivity, although variations in cancer detection rates can be explained not only by reader’s performance but also variations in breast cancer prevalence (Theberge et al. 2014).

Results

Performance Indicators and Quality Assurance Practices

The table in Appendix 6 presents characteristics of breast cancer screening programs in North America, Australia and Europe, and describes quality assurance practices adopted by these programs. When populating this table, data collected from grey literature were supplemented with information from peer reviewed publications. This section contains a summary of data presented in Appendix 6. For a listing of references, see the Appendix.

Centrally organized screening programs are operating in all jurisdictions described in this table except the USA. In the USA, screening is performed opportunistically. Women can self-refer for breast cancer screening in response to recommendations made by their health care providers or based on a possible increased risk of breast cancer.

In the USA, different professional organizations, such as the American College of Radiology, American Cancer Society, the U.S. Preventive Services Task Force, the American Congress of Obstetricians and Gynecologists, issued their recommendations regarding the operation of breast cancer screening.

[3] Definition of specificity from NCI Dictionary of Cancer Terms: “When referring to a medical test, specificity refers to the percentage of people who test negative for a specific disease among a group of people who do not have the disease. No test is 100% specific because some people who do not have the disease will test positive for it (false positive).” https://www.cancer.gov/publications/dictionaries/cancer-terms/def/specificity

Generally, the starting age for screening recommended by these organizations (40-45 years) is younger than the starting age recommended in other jurisdictions (50 years).

RSI conducted a visual examination of identified data on performance indicators of breast cancer screening programs in different jurisdictions. The results were as follows:

- Recall rates in Canada and the USA are higher than those in the European countries or in Australia. At the same time, cancer detection rates in Canada and the USA are comparable to those in Europe and lower than those in Australia.
- Recall rates in France are higher than those in other European countries. Cancer detection rates in the French breast cancer screening program are also higher than cancer detection rates in other European programs.
- In Canada, recall rates increased between 2003 and 2014, and there was no parallel increase in the rates of cancer detection.
- In the US, based on data from Breast Cancer Surveillance Consortium (BCSC), the recall rates and the positive predictive values were relatively stable over time. Note that, unlike in other countries, the USA data represent combined rates for the first and subsequent screens.
- In Australia, recall rates for the first screens increased between 2003 and 2015 in parallel with increasing rates of cancer detection; recall rates for subsequent screens were stable over this period.

There are differences in quality assurance practices between breast cancer screening programs with high recall rates (Canada and USA) and programs with lower recall rates (Europe and Australia).

- In Australia, each Service has a **Designated Radiologist** responsible for all aspects of quality assurance in screen reading. Screen readers receive quarterly Quality Assurance (QA) reports. The QA report includes the reader's recall to assessment rate. The QA report is also provided to the Designated Radiologist and to the Clinical Director of the Service. The Designated Radiologist discusses the reader's QA report and recommends action if required.
- In Australia, quarterly reporting of individual screen reader's performance is practiced. The quarterly Quality Assurance (QA) report includes the reader's recall to assessment rate. The QA report is provided to the reader, to the Designated Radiologist and the Clinical Director of the Service. The Designated Radiologist discusses the reader's QA report and recommends action if required. The Personal Performance in Mammographic Screening (PERFORMS) was implemented by the National Health Service Breast Screening Programme (NHSBSP) in the UK in 1991 and the Breastscreen REader Assessment STrategy (BREAST), a novel **web-based software**, was introduced in Australia in 2011. Screen readers can interpret a standard set of mammograms and receive immediate feedback on their performance.
- In Australia and Europe, **double reading of mammograms is an accepted practice. Discordant opinions are resolved either by consensus, or by arbitration.** Arbitration is undertaken by a radiologist or a panel of radiologists with a high level of expertise in screen reading. In France, where recall rates are higher than in other European countries (and are paralleled with higher cancer detection), the mammogram is read by a second radiologist only if no abnormality is detected by the first reader; when an anomaly is detected, the woman is recalled for further

examination by the first radiologist. In Canada and in the USA, double reading is not a standard practice.

- The **minimum reading volume** by a radiologist required by the European guidelines is 5,000 mammograms per year. The Australian guidelines require reading of at least 2,000 mammograms per year. In France, the first readers must perform at least 500 mammograms per year, and radiologists acting as second readers must interpret at least 1,500 mammograms per year. In the US, a radiologist is required to interpret at least 960 mammographic examinations during a 24-month period. In Canada, a radiologist is recommended to interpret/second read a preferred minimum of 1,000 mammograms per year; however, a minimum of 480 reads per year is still accepted.
- The European guidelines establish slightly **lower acceptable level for recall rates** in initial screens (<7%) as compared to Australia or Canada (<10%). The European guidelines set a desirable level for recall rates: <5% (initial screening); <3% (subsequent screening). In the USA, the national average recall rate (BCSC data), serves as a benchmark for recall rates (e.g., 11.5% based on data through 2013); the Agency for Healthcare Research and Quality (AHRQ) desirable goal for recall rate in screening mammography is <10% (as cited in **Miglioretti et al. 2009**).
- It is recommended that previous mammograms be available to readers at the time of screen reading because comparisons with prior images reduces the likelihood of false-positive findings (**Williams et al. 2015**). This practice is adopted in most breast cancer screening programs. In the USA, comparisons with previous mammograms can be made consistently only if a woman returns to the same provider for subsequent screening rounds (**Williams et al. 2015**).
- **Accreditation** is a quality assurance practice that is common to breast cancer screening services in Australia, Europe and North America. In Australia and the US, accreditation is mandatory while in Europe and Canada it is voluntary.
- In Australia's BreastScreen program, a **National Quality Improvement Plan** for 2018-2020 has been adopted (**BreastScreen Australia, 2018**). It has been acknowledged that "there are standards and targets that are historical that may benefit from review". Specifically, "the recall to assessment target for first screens is often unmet". Establishing a national performance benchmarking program is listed as one of the national priorities.

Mammography Technology

Findings from studies investigating the potential influence of mammography technology are summarized below. More information on review articles and on publications reporting on original research results can be found in Appendices 3 and 5, respectively. For comparison purposes between studies, results discussed in this section of the report were considered statistically significant at a p-value less than 0.05. If the p-value was not reported, the results were interpreted without mention of statistical significance.

Screen-film vs. Digital Mammography

Summary of Review Articles

Earlier systematic reviews (Irwig et al. 2004, Elmore et al. 2005) included few studies comparing screen-film mammography to digital mammography; therefore, no clear conclusion regarding the recall rates could be made. As well, these reviews reported performances that were comparable between the two technologies in terms of cancer detection rates. Later systematic reviews (Vinnicombe et al. 2009, Iared et al. 2011) included more studies and observed considerable heterogeneity in recall rates, with some studies showing significantly lower and others significantly higher recall rates for digital compared to screen-film mammography. Due to significant heterogeneity, Vinnicombe et al. 2009 could not calculate the pooled estimate for the recall rate. In contrast, the pooled estimate was obtained by Iared et al. 2011 (RR = 1.07; 95% CI = 0.94-1.22), and suggests no significant difference in recall rates between the two screening modalities. The results for cancer detection rate were more homogeneous. The pooled estimate obtained by Iared et al. 2011 indicates a significantly better performance of digital mammography in terms of cancer detection: the average relative-risk for cancer detection among patients who underwent digital mammography was 1.17 (95% CI: 1.06-1.29) in relation to screen-film mammography. The findings by Vinnicombe et al. 2009 were also consistent with digital mammography having a higher cancer detection rate than screen-film mammography, but the difference was not statistically significant. Specifically, the pooled estimate for the difference in cancer detection rates between full field digital mammography (FFDM) and screen-film mammography (SFM) was 0.04 (95% CI: -0.03, 0.11) per 100 screening mammograms, which was equivalent to FFDM detecting an extra four cases of breast cancer per 10 000 screening mammograms.

Mohd Norsuddin et al. (2015) [narrative review] concluded that trials comparing digital and film mammography in a screening context demonstrated conflicting results with regards to recall rates. The subject of Le et al. (2016) was the rate of false positives (FP) rather than recall rates. The authors of this review suggests that the increase in the FP rate following the transition to digital mammography demonstrated in some studies, was associated with the use of computer-aided detection (CAD) image interpretation rather than with factors inherent to the digital mammographic image acquisition itself.

Summary of Original Studies

A total of 13 studies comparing performance measures, such as recall rates, cancer detection rates, or positive predictive values, between digital mammography (DM) and screen-film mammography (SFM) were included in this report. The publication dates of studies included were quite recent, and ranged from 2009 to 2018. Study/screening program locations varied greatly, and included Canada (Theberge et al., 2016), Italy (Campari et al., 2016), Spain (Sala et al., 2015), Ireland (Hambly et al., 2009), Norway (Hofvind et al., 2014), Belgium (Van Ongeval et al., 2010), USA (Glynn et al., 2011; Vernacchia et al., 2009), the Netherlands (Karssemeijer et al., 2009; Sankatsking et al., 2018; de Munck et al., 2016; van Luijt et al., 2013), and the United Kingdom (Vinnicombe et al., 2009).

The study by **Campari et al. (2016)** was conducted in Italy and reported on the Reggio Emilia Breast Cancer Screening Program which targeted women 45 to 74 years of age and provided screening either annually (aged 45 to 49 years) or biennially (aged 50 to 74 years) depending on the age group, where double reading with arbitration was used. Results adjusted for age and screening round were reported. A significantly higher adjusted recall rate was observed with DM compared to SFM (RR: 1.46; 95% CI: 1.37, 1.56). However, the study authors observed a decrease in this performance measure with DM where it became similar to the recall rate of SFM following 12 months. A significantly lower PPV was found for DM than SFM (RR: 0.70; 95% CI: 0.59, 0.84). Although a slightly lower adjusted detection rate was observed for DM compared to SFM, the difference was not statistically significant. No significant difference in the DCIS detection rate was observed between the two technologies (RR: 0.91; 95% CI: 0.59, 1.40).

Glynn et al. (2011) reported on a retrospective audit conducted in the US, where performance measures were assessed from 2004 to 2005 for SFM, and in 2007, 2008, and 2009 for DM, among women with a median age of 52 years (range: 27 to 92 years). The recall rate (%) was significantly higher among the time periods of DM use compared to SFM use (SFM: 6.0 [95% CI: 5.7, 6.3]; DM yr1: 7.1 [95% CI: 6.6, 7.6]; DM yr2: 8.0 [95% CI: 7.4, 8.7]; DM yr3: 8.5 [95% CI: 8.1, 9.0]; all p= <0.0001). Although higher cancer detection rates per 1,000 women were also observed among DM compared to SFM (3.34; 95% CI: 2.75, 4.03), results were only statistically significant for the years 2007 (5.28; 95% CI: 4.03, 6.80; p= 0.0061) and 2008 (5.93; 95% CI: 4.36, 7.89; p=0.0016). No significant differences in the PPV_1 were observed between DM and SFM.

They study by **Karssemeijer et al. (2009)** was conducted in the Netherlands and reported on a population-based breast cancer screening program at the Preventicon screening centre which targeted women 50 to 75 years of age, and provided screening every two years. The factor of study was full-field digital mammography (FFDM) with computer-aided diagnosis (CAD) compared to screen-film mammography (SFM); independent double reading with consensus was implemented as the reading approach. The recall rate (%) was significantly higher with FFDM compared to SFM in both initial (4.41 vs. 2.32; p= <0.001) and subsequent screens (1.70 vs. 1.17; p= <0.001). No significant differences were observed between FFDM and SFM for cancer detection rate and invasive cancer detection rate. Significantly higher ductal carcinoma in situ (DCIS) rates (%) were observed for FFDM during both initial (0.22 vs. 0.12; p= 0.015) and subsequent screens (0.12 vs. 0.08; p= 0.007). As well, the PPV of recall (%) were consistently higher for SFM (initial: 26.8; subsequent: 43.1) compared to FFDM (initial: 17.4; subsequent: 30.4).

Vernacchia et al. (2009) reported on a small community-based radiology practice in the US, where performance measures were assessed for SFM in audit 1, and DM in audit 2 to 4. The recall rate among the BI-RADS category of 0 was significantly higher during audit periods of DM compare to SFM (Audit 1: 5.9; Audit 2: 10.2; Audit 3: 7.5; Audit 4: 9.0; p= <0.001). Although the cancer detection rate per 1,000 women screened were significantly higher during audit 2 compared to audit 1 (7.9 vs. 4.1; p= 0.012), no significant differences were observed with audits 3 (5.1; p= 0.42) and 4 (6.9; p= 0.052).

The study by **Sala et al. (2015)** was conducted in Spain and reported on a population-based breast cancer screening program in Barcelona which targeted women 50 to 69 years of age, and provided screening every 2 years, where double reading with arbitration was implemented as the reading approach. Although the recall rate (%) was higher for FFDM relative to SFM during the initial screen (11.73 vs. 11.00; p= 0.032), opposite results were observed during the successive screen (2.50 vs. 3.72; p= <0.001); results for both screening rounds were statistically significant. The cancer detection rate (%) was significantly higher for

FFDM compared to SFM during the initial screen (0.55 vs. 0.39; p= 0.024). However, no significant difference in this performance measure was observed between the two technologies during the successive screen. In the analyses adjusted for radiology unit, age, screening round of diagnosis, no significant difference was observed for screen-detected cancers with FFDM at periods 1 to 4, relative to SFM at period 1. Similarly, the difference in invasive carcinomas detection rates between FFDM and SFM did not reach statistical significance, and this was reflected in the results from the adjusted analysis where no significant difference was observed for the detection of invasive cancers with FFDM at periods 1 to 4, relative to SFM at period 1. The in situ carcinoma detection rate (%) was only significantly different between FFDM and SFM during the initial screen (FFDM vs. SFM: 0.12 vs. 0.06; p= 0.031). In the adjusted analysis, a significantly higher DCIS detection was observed with FFDM at periods 2 to 4, relative to SFM at period 1. The positive predictive value (%) was consistently higher with FFDM compared to SFM in both initial (6.43 vs. 4.20; p= 0.010) and successive (14.64 vs. 11.14; p= 0.004) screens; both results were statistically significant.

Hambly et al. (2009) was conducted in Ireland, and reported on the Irish National Breast Screening Program (INBSP) which targeted women 50-64 years of age, and provided screening every 2 years, where unblinded double reading with consensus was used. The recall rate (%) was significantly higher with FFDM compared to SFM in both first (7.3 vs. 5.7; p= <0.001) and subsequent (2.8 vs. 2.0; p= <0.001) screens. Although the cancer detection rate per 1,000 screens were significantly higher with FFDM relative to SFM during the subsequent screens (5.7 vs. 4.4; p= 0.008), differences between the two technologies were not statistically significant during the first screen. Similarly, the invasive cancer detection rate (FFDM vs. SFM: 4.4 vs. 3.6; p= 0.047) and DCIS rate (FFDM vs. SFM: 1.2 vs. 0.8; p= 0.036) per 1,000 screens were only statistically significant in the subsequent screen. No significant differences in PPV_1 were observed between FFDM and SFM.

The study by **Sankatsing et al. (2018)** was conducted in the Netherlands, and reported on the Dutch breast cancer screening programme (BCSP) which targeted women 50 to 74 years of age, and provided screening biennially, where double reading with either consensus or arbitration was implemented as the reading approach. Higher recall rates per 1,000 screens were observed with DM compared to SFM (21.0 [95% CI: 20.8, 21.2] vs. 16.0 [95% CI: 15.9, 16.1]). As well, the detection rates for all cancers (6.2 [95% CI: 6.1, 6.3] vs. 5.4 [95% CI: 5.3, 5.4]), DCIS (1.1 [95% CI: 1.1, 1.2] vs. 0.83 [95% CI: 0.81, 0.86], and invasive cancers (5.1 [95% CI: 5.0, 5.2] vs. 4.5 [95% CI: 4.5, 4.6]) per 1,000 screens were higher with DM than SFM. However, the positive predictive value (%) was greater for SFM compared to DM (34.9 [95% CI: 34.5, 35.2] vs. 31.5 [95% CI: 31.1, 31.9]). All results were adjusted for age.

De Munck et al. (2016) was conducted in the North-Netherlands, and reported on the Dutch breast cancer screening program which targeted women 50 to 75 years of age, and provided screening biennially, where independent double reading with consensus was used. Although a significantly higher proportion of women were recalled (%) during the initial screen with FFDM compared to SFM (3.02 vs. 2.07; p= <0.001), results were not statistically significant during the subsequent screens. No significant differences were observed with screen detected breast cancers, DCIS, or invasive cancers per 1,000 screened women between the two mammography technologies. Additionally, the positive predictive value for screen detected breast cancers (%) was significantly greater for SFM compared to FFDM during the initial screen (25.6 vs. 19.9; p= 0.002); the difference in this performance measure between SFM and FFDM were not statistically significant in subsequent screens.

The study by **Theberge et al. (2016)** was conducted in Canada, and reported on the Quebec Breast Cancer Screening Program (Programme Québécois de Dépistage du Canada due Sein [PQDCS]) which targeted women 50 to 69 years of age, and provided screening biennially, where single reading was implemented as the reading approach. Performance measures were compared between SFM, computer radiography (CR), and digital direct radiography (DR), as well as by manufacturers of CR. Significantly higher recall rates were observed for CR (OR: 1.03; 95% CI: 1.01, 1.06) and DR (OR: 1.25; 95% CI: 1.19, 1.30) compared to SFM (OR: 1.00). When assessing CR by manufacturer relative to SFM, recall rates were significantly higher for CR-Fuji (OR: 1.05; 95% CI: 1.02, 1.07), significantly lower for CR-Agfa (OR: 0.93; 95% CI: 0.89, 0.98), and not significantly different for CR-Kodak (OR: 1.02; 95% CI: 0.97, 1.08). No significant differences in the detection rate, invasive detection rate, or DCIS detection rate were observed between CR and DR compared to SFM. The PPV was significantly lower for CR-Kodak (OR: 0.81; 95% CI: 0.67, 0.98) relative to SFM (OR: 1.00); all other comparisons yielded results that were not significantly different. Refer to the tabular summaries for the list of adjusted covariates.

Vinnicombe et al. (2009) was conducted in the United Kingdom, and reported on the Central and East London Breast Screening Service (CELBSS) which targeted women aged 50 years or above, and provided screening every 3 years, where unblinded double reading with arbitration was used. No significant differences were observed between SFM and FFDM for cancer detection rates, recall rates, and PPV, even after stratification by age groups (≤ 60 years vs. >60 years); these results were adjusted for age, ethnicity, referral type, and area of residence.

The study by **Hofvind et al. (2014)** was conducted in Norway, and reported on the Norwegian Breast Cancer Screening Program (NBCSP) which targeted women 50 to 69 years of age, and provided screening every 2 years. The reading approach implemented was independent double reading with the use of a consensus or an arbitration meeting when screening mammograms were scored a 2 (“probably benign”), 3 (“intermediate”), 4 (“probably malignant”), or 5 (“high suspicion of malignancy”) by either or both radiologists. The overall recall for further assessment (0.34 vs. 0.29; p= <0.001), total screening-detected cancer (0.56 vs. 0.52; p= 0.005), and screen-detected invasive breast cancer (0.47 vs. 0.42; p= <0.001) per 1,000 examinations were significantly higher for SFM compared to FFDM. In contrast, the overall screen-detected cancer for DCIS per 1,000 examinations were significantly higher for FFDM than SFM (0.11 vs. 0.09; p= 0.019). In addition, SFM exhibited a significantly higher PPV (%) during baseline examinations (12.9 vs. 10.0; p= <0.05); however, during subsequent examinations, a significant effect between SFM and FFDM was observed in the opposite direction (SFM after SFM: 19.3; FFDM after SFM: 21.63; FFDM after FFDM: 22.73; p= <0.05). The study also performed an analysis adjusted for screening modality, period, and age; the incidence rate ratio (IRR) for screening-detected DICS was 1.43 (95% CI: 1.20, 1.71) for FFDM after SFM and 1.32 (1.07, 1.64) for FFDM after FFDM relative to SFM after SFM (IRR: 1.00); no significant differences between FFDM after SFM and FFDM after FFDM were observed compared to SFM after SFM for screening detected breast cancer and invasive breast cancer.

Van Ongeval et al. (2010) was conducted in Belgium, and reported on a decentralized screening program which targeted women 50 to 69 years of age, and provided screening biennially. The reading approach implemented was independent double reading with the use of a third reader when mammograms were scored a 3 (“probably benign finding”), 4 (“probably malignant finding”), or 5 (“malignant finding”) by either reader. Compared to the first SFM control population, which included SFM use among three regional screening units that were the first to change to DM, the recall rate (%) was significantly higher with SFM compared to DM during subsequent screens (1.58 vs. 1.20; p= 0.03). In contrast, results were not statistically significant during the initial screen. No significant results were reported for cancer detection rate (%) or PPV (%). However, significantly greater DCIS (%) was detected with SFM than DM (0.16 vs. 0.07; p= 0.02). In comparison with the second SFM control population, which included SFM indicators of 47 mammographic units, recall rate, cancer detection rate, and PPV were not statistically different between the two mammography technologies.

The study by **van Luijt et al. (2013)** was conducted in the Netherlands, and reported on the Dutch national breast cancer screening programme, which targeted women 50 to 75 years of age, and provided screening biennially. The technology groups investigated included the following: DM (“DM read by a reading unit reading both SFM and DM”), SFM (“SFM ready by a reading unit reading both SFM and DM”), and SFM Only (“SFM read by a reading unit reading only SFM”). Significantly higher recall rates in percentages (DM: 2.0; SFM: 1.6; SFM Only: 1.6; p= <0.001) and detection rates per 1,000 screens (DM: 5.9; SFM: 5.1; SFM Only: 5.0; p= <0.001) were reported for DM compared to SFM. In contrast, the PPV (%) was significantly higher for SFM relative to DM (DM: 31.2; SFM: 34.4; SFM Only: 34.2; p= <0.001).

<u>*Overall Summary*</u>

The review articles suggest that there is no clear evidence of one technology being superior to the other in terms of recall rates. The performance of digital mammography in terms of cancer detection rates is better than (or at least comparable to) that of screen-film mammography. Among articles reporting on original research, evidence on the recall rate, cancer detection rate, and PPV of DM relative to SFM was unclear as some inconsistencies were observed across and within studies (i.e. differences between initial and subsequent screens).

Computer-Aided Detection Systems

Summary of Review Articles

Systematic Reviews on Single Reading with CAD vs. Single Reading: Earlier systematic reviews (Irwig et al. 2004; Elmore et al. 2004) included few studies and provided no clear evidence of the effect of CAD on recall rates or cancer detection rates. The review by Taylor and Potts (2008) found an increase in recall rate with CAD among all studies; however, significant heterogeneity was also observed. The overall pooled estimate for the effect of CAD on recall rates was statistically significant (OR=1.10; 95% Cl: 1.09, 1.12); these results remained statistically significant when stratified by study design (matched and unmatched studies). As well, the authors found no clear evidence of an effect of CAD on cancer detection rates. Noble et al. (2009) focused on FP rates and concluded that CAD increases the recall of healthy women.

Systematic Reviews on Double Reading vs. Single Reading with CAD: Irwig et al. (2004) identified only one study that reported on incremental true and false positives with single reading and CAD as compared to double reading. The authors of the review found that the increments seen in this study were difficult to quantify. Conclusions regarding the differences in false and true positives between the two groups being compared were not clear in this review. Azavedo et al. (2012) identified four studies, and only one of the four was of moderate quality. This moderate quality study reported significantly higher recall rates associated with single reading plus CAD compared to double reading. However, no statistically significant differences in cancer detection rates were observed between the two groups. The results of the other three studies were inconsistent. The authors of this review concluded that the evidence was insufficient to make conclusions regarding the accuracy of single reading with CAD compared to double reading.

Narrative Reviews: Studies reviewed by Houssami et al. (2009) showed that CAD can improve cancer detection rates of a single reader and at the same time substantially increase the recall rate. The authors also concluded that CAD does not perform as well as double reading in organized breast cancer screening programs where double reading is the standard of care. Houssami et al. (2009) noted that manufacturers of CAD systems worked on improvements of CAD algorithms to increase specificity by reducing false prompts. Astley and Gilbert (2004) indicated that the introduction of CAD may have a different effect on the reader's performance depending on the type and level of the reader's experience.

Summary of Original Studies

A total of 6 studies investigating the influence of CAD on performance measures of interest were included in this report. The publication dates ranged from 2009 to 2018, and study/program locations varied between Barcelona (Bargallo et al., 2014), Spain (Sanchez Gomez et al., 2011), the US (Fenton et al., 2011; Lehman et al., 2016; Gromet et al., 2018), and the UK (James and Cornford, 2009).

The study by **Bargallo et al. (2014)** was conducted in Barcelona, and reported on a population-based breast cancer screening program in Sants-Montjuic, Les Corts, and Eixample Esquerre, which targeted women 50 to 69 years of age, and provided screening every 2 years. This study compares blinded double reading with arbitration to single reading with CAD. A greater recall (7.02% vs. 3.94%) and cancer detection rate (6.10% vs. 5.25%) was observed among single reading with CAD. However, the positive predictive value of recall was greater among double reading with arbitration (13.32% vs. 8.69%).

Sanchez Gomez et al. (2011) was conducted in Spain, and reported on a population-based breast cancer screening program which included women 45 to 65 years of age, and provided screening biennially. The current study compares performance measures of single reading to those of single reading with CAD. No statistically significant difference was observed with the recall rate (Pre-CAD: 7.2% vs. Post-CAD: 7.6%). The detection rate per 1,000 women was significantly higher with CAD than without (Pre-CAD: 4.3‰ vs. Post-CAD: 4.4‰; p= <0.005). A slight difference in the PPV measure between the radiologist (6.4%) compared to the radiologist with CAD (6.1%) was also observed.

The study by **Fenton et al. (2011)** was conducted in the US, and reported on the Breast Cancer Surveillance Consortium (BCSC) from January 1, 1998 to December 31, 2006 which included women 40 years of age or older. CAD was used in 25 of 90 BCSC facilities, and comparisons were made between performance measures of SFM and SFM with CAD. Among facilities using CAD, unadjusted results observed after CAD implementation demonstrated a significant increase in recall rate (after CAD: 8.9% [95% CI: 8.8, 9.0] vs. before CAD: 8.4% [95% CI: 8.3, 8.5]; p= <0.001), a significant decrease in the positive predictive value (after CAD: 3.6% [95% CI: 3.4, 3.9] vs. before CAD: 4.3% [95% CI: 4.1, 4.5]; p= <0.001), as well as a significant decrease in the detection rate of all breast cancers (after CAD: 3.2 [95% CI: 3.0, 3.5] vs. before CAD: 3.6 [95% CI: 3.4, 3.8]; p= 0.01) and invasive breast cancers (after CAD: 2.3 [95% CI: 2.1, 2.5] vs. before CAD: 2.8 [95% CI: 2.7, 3.0]; p= <0.001) per 1,000 mammograms. Although the unadjusted DCIS detection rate in percentages were significantly higher after CAD implementation (after CAD: 24.9% vs. before CAD: 20.0%; p= 0.003), no significant difference was observed when investigating the DCIS detection rate per 1,000 mammograms. Odds ratios (OR) were acquired between CAD use compared to non-CAD use, and were adjusted for mammography registry, age, breast density, time since prior mammography, hormone replacement therapy, and examination year. In these analyses, a significantly lower PPV_1 was observed with CAD use (OR: 0.89; 95% CI: 0.80, 0.99; p= 0.03); however, adjusted results were not significant for the detection rates of all breast cancers, invasive breast cancers, and DCIS when comparing CAD use to non-CAD use.

A study by **Lehman et al. (2016)** was conducted in the US, and reported on the Breast Cancer Surveillance Consortium (BCSC) from January 1, 2003 to December 31, 2009 which included women 40 to 89 years of age. Differences in performance measures were investigated between two technology groups of interest: DM vs. DM with CAD. Odds ratios comparing DM to DM with CAD were adjusted for site, age group, race/ethnicity, time since last mammogram, and calendar year of exam. The differences in recall, total cancer detection, and invasive cancer detection rates were not statistically different between the two technology groups. In contrast, the DCIS detection rate per 1,000 exams was found to be significantly higher for DM with CAD than DM alone (CAD: 1.19 [95% CI: 1.0, 1.3] vs. No CAD: 0.95 [95% CI: 0.7, 1.2]; p= <0.03); these results were also reflected in the adjusted odds ratio of 1.39 (95% CI: 1.03, 1.87; p= 0.031).

The study by **Gromet et al. (2008)** was conducted in the US, and reported on a community-based mammography program in Charlotte, NC. In this analysis, performance measures of single reading with CAD were compared to those of double reading, and the first reader in a double-reading program. A significantly higher recall rate was observed with double reading compared to single reading with CAD (single reading with CAD: 10.6% vs. double reading: 11.9%; p= <0.0001); however, differences for the detection rate (per 1,000 patients), and PPV_1 were not statistically significant. In comparison to the performance measures of the first reader, a significantly higher recall rate was observed for single reading with CAD (single reading with CAD: 10.6% vs. first reader: 10.2%; p= <0.0001); similar to the other comparison group of interest, no significant differences were observed for detection rate and PPV_1.

James and Cornford (2009) investigated the difference in performance between the use of CAD compared to a third reader as an arbitrator for cases of discordant double-reading. Following arbitration, the proportion of women recalled was greater with the use of CAD than a third reader (68% vs. 47%); this recall of an additional 50 women could possibly result in a relative increase in the overall recall rate by 10% (3.1% to 3.4%). However, the use of CAD as an arbitrator resulted in a significantly higher proportion of normal women recalled (66.8% vs. 43.9%; p= <0.001). In regard to the cancer detection among arbitrated lesions, only one additional cancer case was detected with CAD.

Overall Summary

The review articles suggest that the introduction of CAD may increase the recall rate as compared to the performance of a single reader; there is no clear evidence of an effect of CAD with single reading on cancer detection rates. The evidence is insufficient to make conclusions regarding the accuracy of single reading with CAD compared to double reading.

Results from original research showed higher recall and cancer detection rates, as well as lower PPV with the use of single reading (SR) with CAD compared to SR alone; however, the status of statistical significance may have varied between studies (Sanchez Gomez et al. 2011; Gromet et al., 2008). Study findings were inconsistent for SR with CAD compared to double reading (DR) for performance measures of interest, including recall rate, cancer detection rate, and PPV (Bargallo et al., 2014; Gromet et al., 2008). In the comparison between SFM and SFM with CAD, a significantly higher unadjusted recall rate was observed with the use of CAD. Significantly lower cancer detection and invasive cancer detection rates were observed with SFM alone; however, these results were no longer statistically significant following the adjustment of covariates. As well, the PPV was found to be significantly higher among SFM compared to SFM with CAD. Although the DCIS detection rate per 1,000 mammograms was not significantly different between groups, this measure as a percentage (%) was significantly greater among SFM with CAD; following adjustment, the OR for the DCIS detection rate was no longer statistically significant (Fenton et al., 2011). No significant differences were observed between DM and DM with CAD for recall, cancer detection, and invasive cancer detection rates. In contrast, the DCIS detection rate was significantly higher for DM with CAD compared to DM alone (Lehman et al., 2016). In the comparison between the use of CAD and a third reader for arbitration in double reading, the proportion of women recalled, normal women recalled, and cancer detected among arbitrated lesions was higher with CAD (James and Cornford, 2009).

Tomosynthesis

Summary of Review Articles

Findings from systematic (**Svahn et al. 2015**; **Hodgson et al. 2016**; **Nelson et al. 2016**; **Pozz et al. 2016**) and narrative reviews (**Cole et al. 2016**; **Gilbert et al. 2016**; **Vedantham et al. 2015**; **Skaane et al. 2017)** suggest that digital breast tomosynthesis (DBT) has the potential to decrease recall rates and increase cancer detection rate.

Summary of Original Studies

Most original studies identified for this report are in line with this conclusion, with occasional opposite results. For example, increased recall rates with DBT were seen by **Friedewald et al. 2014** in two of thirteen breast cancer screening sites included in their multicenter study. One study (**Giess et al. 2017**) demonstrated that, in the overall cohort, there was no significant difference in recall rates between DBT and full field digital mammography (FFDM); however, the recall rate at baseline screening examination was lower with DBT. In this study, the cancer detection rate and the positive predictive value of recall were higher with DBT than with FFDM. **Lourenco et al. 2015** found that, although overall recall rate was lower with DBT than with two-dimensional digital mammography (DM), DBT resulted in more recalls for some specific types of breast abnormalities, specifically for masses, distortions, and calcifications. Benefits of DBT may vary by patient's characteristics such as age and breast density (**Sharpe et al. 2016**). **Giess et al. 2017** pointed out that the introduction of DBT might not produce a straightforward effect on recall rates because not only technology but also other factors, such as patients' characteristics, radiologist's experience, training, and tolerance for errors, affected recall rates.

Overall Summary

In summary, the implementation of DBT was associated with reduced recall rates in most studies. DBT may be associated with increased recall for some specific breast abnormalities. Most studies show increased cancer detection rates with DBT. Benefits of DBT may vary by patient's characteristics such as age and breast density.

Synthesized Mammography

Because synthesized mammography was not initially included in the scope of this review, RSI did not conduct an in-depth analysis of literature on recall and cancer detection rates associated with this technology. The following data were obtained from study abstracts only; full texts of these articles were not reviewed, and data were not abstracted in the data abstraction tables.

Summary of Review Articles

Synthesized mammography (SM) images are designed for interpretation with digital breast tomosynthesis (DBT) as a complement. Examination using a combination of two-dimensional full-field digital mammography (DM) with DBT provides performance advantages over DM alone, and at the same time approximately double the radiation dose. Implementation of synthesized mammography decreases the radiation dose nearly two-fold while providing the advantages of DBT. The technology was approved by the United States Food and Drug Administration (FDA) in 2013 (**Durand, 2018**).

Summary of Original Studies

Durand (2018) [narrative review] concluded that the available studies "substantiate the claim that the benefits of low recall rates seen with digital breast tomosynthesis implementation can be upheld with synthesized mammography, while simultaneously decreasing radiation dose." The author also concluded that DBT practice patterns is expected to change in favor of increasing implementation of synthesized mammography as a replacement for conventional two-dimensional mammography in conjunction with DBT.

Based on RSI's overview of the abstracts, it appears that two types of comparisons were performed: 1) the performance of digital breast tomosynthesis in combination with synthesized mammography (DBT+SM) was compared with that of (full-field) digital mammography (DM) alone; 2) the performance of DBT in combination with SM (DBT+SM) was compared with that of a combination of DBT with DM (DBT+DM).

DBT+SM vs. DM

One study conducted in the USA (Aujero et al. 2017) demonstrated significantly lower recall rates with DBT+SM (4.3%) vs. DM alone (8.7%), and no significant differences in cancer detection rates between the two modules. In contrast, Bernardi et al. (2016) [Italy] demonstrated significantly higher false positive recall rates (4.45% vs. 3.42%)[4] and significantly higher cancer detection rates 8.8 per 1000 vs. 6·3 per 1000) for DBT+SM compared to DM. Caumo et al. (2017) [Italy] found that overall recall rates were similar for the two screening modules (relative risk (RR), 0.95; 95% CI: 0.84, 1.06) while the recall rate with invasive assessment was higher for DBT+SM than for DM (RR, 1.93; 95% CI: 1.31, 2.03); cancer detection rates were significantly higher in patients screened with DBT+SM than in those screened with DM.

DBT+SM vs DBT+ DM

Three studies conducted in the US (Ambinder et al. 2018; Aujero et al. 2017; Zuckerman et al. 2018) demonstrated that recall rates were significantly lower with DBT+SM as compared to DBT+DM; there were no differences in cancer detection. One Italian study (Bernardi et al. 2016] found no appreciable differences in false positive recalls (4.45% for DBT+SM vs. 3.97% for DBT+DM) or cancer detection rates (8·8 per 1000 for DBT+SM vs. 8·5 per 1000 for DBT+DM] between the two modules.

Overall Summary

In summary, implementation of synthesized mammography in conjunction with digital breast tomosynthesis may be beneficial: synthesized mammography appears to preserve the performance benefits provided by DBT, and at the same time reduces the dose of radiation.

Quality Assurance Practices

Reading Volume

Summary of Review Articles

It appears reasonable to suggest that reading volume has some influence on the accuracy of mammogram interpretation. Readers who read a larger number of mammograms may have lower recall and false positive results because they acquire a better knowledge of normal presentations seen on screening mammograms at an earlier stage in their career (**Coldman et al. 2006**; **Mohd Norsuddin et al. 2015**; **Le et al. 2016**). It has been suggested that differences in the minimum required number of mammograms to be interpreted per year by a radiologist contribute to the differences in recall and false positive rates between North America and other jurisdictions such as the UK (**Le et al. 2016**). However, in contrast to these expectations, reviews of relevant data by experts in the field (e.g., **Coldman et al. 2006; Buist et al. 2011**; **Theberge et al. 2014**)[5] found that studies had shown widely varying results regarding the relationship between reading volume and recall/false positive rates. As discussed by **Buist et al. 2011,** in some cases conflicting results had been obtained by different groups of investigators using data from overlapping populations. Peer-reviewed publications identified by RSI for this report also show inconsistencies, with

[4] Durand (2018) believed that the increase with synthesized mammographu "may have occurred because all recalls were included, after reading any part of the exam, even s2D [synthesized mammography] alone… As stated earlier, in clinical use, s2D is always read in conjunction with the DBT component."

[5] These three articles report on the results of original research as well as provide summaries of relevant information from other published studies.

results varying from positive, inverse, and no association between reading volume and recall/false-positive rates. These studies are summarized below, and a table with more details can be found in Appendix 5.

Summary of Original Studies

Studies in Europe and the UK

Alberdi et al. 2011 observed a decreasing trend in the risk of overall false positive results and in the risk of false positive results leading to an invasive procedure in Spanish population-based breast cancer screening programs. This trend was observed in reading volumes ranging from less than 500 to more than 15,000 mammograms per year.

In the study by **Cornford et al. (2011)** [East Midlands Breast Screening Programme, UK], screen readers who interpreted less than 15,000 mammograms over 3 years (i.e., on average <5,000 mammograms per year) were considered as "low volume readers". Although the results from this analysis demonstrated that "low volume readers" had the highest median recall rate compared to the other two reading volume groups combined (20,000 to <25,000 and ≥25,000 mammograms over 3 years), this difference was not statistically significant. As well, the "low volume readers" did not differ significantly from the other groups in terms of cancer detection. **Duncan and Scott (2011)** used similar cut-off points to group reading volumes of screen readers practicing in Scotland and found no significant group difference in recall rates or sensitivity. Both studies included screen readers of different qualifications (radiologists, radiographers, and breast clinicians) and did not account for these differences. Also, the analyses were not adjusted for other radiologists' or patients' characteristics, and this may possibly be due to the small sample sizes.

Studies in the USA

All studies conducted in the USA and identified for this report are based on data from the mammography registries participating in the Breast Cancer Surveillance Consortium (BCSC).

Analyses by **Buist et al. 2011** included annual total (up to ≥5000 mammograms/year), screening (up to ≥5000 mammograms/year) and diagnostic (up to ≥1000 mammograms/year) reading volumes. The percent of time spent in screening ("screening focus") was another factor of interest. The authors concluded that "higher interpretive volume was associated with clinically and statistically important lower rates of false-positive results and numbers of women recalled per cancer detected - without a corresponding decrease in sensitivity or CDR [cancer detection rate]". Radiologists with a greater screening focus had significantly lower false-positive rates and CDR. CDR was also lower for radiologists with low diagnostic volumes.

Smith-Bindman et al. (2005) included in their analysis radiologists who interpreted greater than 480 mammograms per year and divided the radiologists into 6 groups by annual interpretive volume from 481-750 (reference group) to greater than 4,000 mammograms. These researchers analyzed specificity, sensitivity and the total accuracy of mammogram interpretation. There was no obvious trend in specificity with increasing reading volume; only one group consisting of radiologists with reading volumes of 2,501-4,000 mammograms per year demonstrated a significant increase in specificity compared to the reference group. Specificity of radiologists who interpreted more than 4,000 mammograms per year was of similar magnitude to that in the reference group. Also, specificity was significantly increased in radiologists who focused on screening mammography (the ratio of screening to diagnostic mammograms >5 vs. <5). There was no trend in sensitivity or overall accuracy with increasing reading volume, and no group had significantly increased sensitivity compared to the reference group. Radiologists who focused on screening mammograms had significantly lower sensitivity and significantly higher overall accuracy.

Elmore et al. (2009) analyzed three groups of radiologists by annual average reading volume: ≤1,000, 1,001-2,000 and >2,000. They found no association between annual volume of mammograms and any of the interpretive performance indicators studied (recall rates, false-positive rates and positive predictive values of recall). These indicators were not significantly different between radiologists whose interpretive volume included 83% or more of screening mammograms and those with less than 83% of screening mammograms.

Barlow et al. 2004, analyzed data from BCSC registries, and demonstrated higher recall rates and sensitivity, as well as lower specificity in radiologists who interpreted more than 1,000 mammograms in the past year as compared to those who interpreted fewer mammograms. The authors concluded it was unlikely that increasing volume requirements would increase mammography performance unless there was an adequate feedback on cancer outcomes and radiologists' discriminative skills.

Canadian studies

Coldman et al. (2006), used data from several Canadian provincial screening programs, and found no consistent effect of the radiologist's reading volume (up to ≥5,000 mammograms per year) on abnormal call rates or cancer detection rates. However, there was a consistent pattern of increasing positive predictive value (PPV) with higher reading volumes. Radiologists who interpreted fewer than 480 mammograms per year were not included in these analyses.

The aim of the study by **Theberge et al. (2005)** was to determine how the caseload of individual radiologists and the number of screenings performed in the facility influenced the rate of false-positive results and cancer detection rates. These investigators analyzed data from the Quebec Breast Cancer Screening Program. The false-positive rate ratio decreased significantly with increasing individual screening volume of radiologists. In contrast, no association was observed between the screening volume of facilities and false-positive rates. The rate of cancer detection was positively associated with the number of screenings performed in the facility; however, no association was observed with the individual screening volume of radiologists. The investigators also analyzed the combined effect of these two factors and concluded that "the overall performance of screening mammography seems to be maximized when screenings are performed in larger centres and when, in these centres, mammograms are read by radiologists who interpret a large volume of films." These analyses were adjusted for important radiologists' and patients' characteristics. The analyses of radiologists' screening volumes were adjusted for facilities' screening volumes and vice versa.

A more recent study with partially overlapping authorship (**Theberge et al. 2014)** found a significant decrease in false positive rates associated with an increase in radiologist's annual interpretive volume. This significant decreasing trend in false positives was seen for total annual volume of mammograms as well as for screening and diagnostic volumes. The reduction in false-positives was greater at the lower volume for all volume types, and the curve stabilized at higher volumes. Sensitivity was not associated with any volume type, and accuracy defined as sensitivity/false-positive rate significantly increased with increasing reading volumes of all types. The authors concluded: "…this study suggests that the minimal volume requirement of 500 mammograms annually adopted in North America is justified. Radiologist accuracy may be compromised when interpretive volume consistently falls short of this minimum requirement. Raising the interpretive volume of radiologists may help to minimize false-positive screens without sacrificing sensitivity. Our results demonstrate that potential gains in accuracy with increases in volume may be greater up to an annual interpretive volume of approximately 3000 mammograms." Because the earlier study (**Theberge et al. 2005**) demonstrated a significant influence of the number of screenings performed in the facility, **Theberge et al 2014** adjusted their analyses for facility's volume and type, as well as for many important radiologists' and patients' characteristics.

Possible reasons for discrepant findings

Investigators identified several possible explanations for the inconsistencies. These include methodological differences, such as different approaches for measuring reading volume (self reported volume which is vulnerable to recall bias vs. observational data), different study designs, performance measures, modeling methods and covariates included in the models (**Coldman et al. 2006**; **Buist et al. 2011**; **Theberge et al. 2014**). As discussed by **Theberge et al. (2014)**, some studies excluded radiologists based on their experience and reading volume while others did not; also, some studies either did not adjust for potential confounders or adjusted for only a few patient's characteristics. Few studies adjust for radiologist's characteristics. At the same time, **Coldman et al. 2006** found that the influence of inter-radiologist variation on the abnormal interpretation rate was one of the strongest in their study examining the relationship between radiologist's reading volume and interpretive performance. Performance of radiologists within the same group defined by reading volume is highly variable, and this variability is largely unexplainable, suggesting that the volume-performance relationship is complex and several factors, such as radiologist's training, years of experience, number of cancers interpreted, screening vs diagnostic volume, may influence it (**Buist et al. 2011**). There may be other factors that influence the relationship, for example whether feedback is provided regarding radiologists' discriminative skills (**Barlow et al. 2004**). Also, reverse causation cannot be ruled out: radiologists who interpret more accurately choose to interpret more mammograms (**Smith-Bindman et al. 2005; Buist et al. 2011**).

Possible reading volume threshold (decline or saturation in performance)

If there is a positive relationship between the reading volume and performance indicators, there may be a threshold over which there is saturation or decline in performance (**Mohd Norsuddin et al. 2015**).

Cornford et al. (2011) used data from the East Midlands Breast Screening Program (UK), and demonstrated that the median cancer-detection rate of mammogram readers interpreting 25,000 or more mammograms over a 3 years period was significantly lower than that of radiologists interpreting fewer mammograms over the same period. The recall rate of these high-volume readers was also significantly lower than the recall rates of those interpreting fewer mammograms. **Cornford et al. (2011)** concluded that there may be an upper limit above which reader performance deteriorates in terms of cancer detection.

Duncan and Scott (2011) performed a study to investigate whether the finding by Cornford et al. [at the time published as a conference abstract] could be replicated in Scotland. Mammogram readers were divided into high, medium or low volume readers using thresholds similar to those used by Cornford et al. The findings by **Duncan and Scott (2011)** did not support the suggestion that reading performance declines with a 3-year volume of 25,000 or more. The authors acknowledge that, because at the time of writing, the study by Cornford et al. was not published, details of their methods were not known, and the two studies may not be directly comparable. Also, as discussed above, both studies included readers of different qualifications and did not account for these differences; other radiologists' or patients' characteristics were not considered in these analyses.

Given-Wilson and Blanks (2011) suggest looking at the trade-off between sensitivity and specificity. Using a plot of the likelihood of women recalled having cancer (PPV of recall) against the percentage of women recalled, the authors show that, "as recall rates fall to very low values there is a point where cancer detection begins to fall off. This is because the reader is simply not recalling enough women to identify those with subtle signs of cancer". **Given-Wilson and Blanks (2011)** believe that the highest volume readers studied by **Cornford et al. (2011**) reached the point where a low recall rate begins to affect sensitivity. At this point, detection rate may be improved by a slight increase in recall rate. **Given-Wilson and Blanks (2011)** also point out that Cornford et al. and Duncan and Scott have looked at individual reader performance while in real-life screening individual errors can be mitigated using double-reading and arbitration or consensus. **Given-Wilson and Blanks (2011)** conclude that "there is some suggestive evidence that high-volume readers need to monitor their recall rates to ensure that high-volume reading does not lead to a lessening in detection rate as a result of too low a recall rate".

Possible key study
RSI suggests that the Partnership consider **Theberge et al. (2014**) as a possible key study examining the relationship between radiologists' reading volume and their interpretive performance. Although the investigators analyzed false positive rates rather than recall rates, the study was conducted in Canada, and many important patients' and radiologists' characteristics, as well as characteristics of mammography facilities, were accounted for in these analyses.

Approaches to double reading

Double reading vs. single reading

Summary of Review Articles
The systematic review and meta-analysis by Taylor and Potts (2008) demonstrated that, compared to single reading, double reading with unilateral recall (when the opinion of only one reader is the basis for recall) or double reading with mixed recall practice, was associated with increased recall rates while double reading with resolution of discordant opinions via consensus or arbitration was associated with reduced recall rates. The authors acknowledge significant heterogeneity among studies and large uncertainty regarding the influence of double reading on recall rates. Cancer detection rates were increased at double reading compared to single reading regardless of recall strategy. There was no evidence of heterogeneity among studies regarding cancer detection rates.

In their systematic review and meta-analysis, Posso et al. (2017) found no significant differences in false positive rates and cancer detection rates between double reading and single reading of digital mammograms. Although recall strategies adopted in each included study are reported in this review article, the analysis was not stratified by recall strategy, most likely because very few studies were included in this review (two reporting on false positives and three reporting on cancer detection rate). The authors acknowledged large uncertainties regarding the effects of double reading on cancer detection and false positives.

Summary of Original Studies

Original publications identified for this report demonstrate that practices of double reading, recall strategies and readers' characteristics differ among mammography programs. As a result, generalizations of study results may be problematic. For example, the two mammogram readers in a double-reading program may or may not be aware of each other's opinions (blinded or non-blinded double reading), patients may be recalled based on the opinion of only one reader (unilateral recall), or a common conclusion is required for recall. The common conclusion can be reached via consensus, arbitration, or their combination. In some studies, arbitration was performed only if the two radiologists could nor reach consensus (**Posso et al. 2016**). Arbitrators may or may not be blinded to the reason for disagreement between the two original readers. Also, arbitration may be undertaken by a single third reader or by an arbitration panel (**Posso et al. 2016**; **Shaw et al. 2009; Taylor-Phillips et al. 2016**). The arbitration panel may or may not include the two original screen readers (**Shaw et al. 2009**).

Unilateral recall

Caumo et al. 2011a [Italy] studied the benefits of so called "delayed second reading procedure as an adjunct to real-time reading with immediate assessment". Women were recalled for further assessment based on first reader's opinion, and the second reading "followed in a separate session". The second reader was aware of the opinion of the primary reader. Recall rate at first reading was 13%, and 2.7% of screened women were recalled based on the opinion of the second reader only (21.2% relative increase compared to the first reading). The second reading was associated with an absolute increase in cancer detection rate of 0.93 per 1000 screened (+13.1% relative to first reading).

Similar results were obtained by **Ciatto et al. 2005** [Italy]. In this double-reading program, the second reader was aware of the first reader's opinion. Referral to assessment was prompted by suspicion by either reader with no consensus or arbitration of discordant cases. Double reading was associated with +0.70% additional referral rate (24% increase relative to single reading) and +0.024% cancer detection rate (6.4% increase relative to single reading). The authors concluded that the contribution of second reading to cancer detection rate was limited in magnitude. Estimated additional costs were 2.70 euros per woman screened with double reading, 11,168 euros per additional cancer detected, and 11,585 euros per cancer detected by single reading.

Resolution of discordant opinions via consensus or arbitration

Roman et al. 2012 [Spain] found that, compared to single reading, double reading was associated with higher false positive rates (OR=2.06; 95% CI: 2.00, 2.13) and cancer detection rates (OR=1.08; 95% CI: 1.04, 1.12). It is unclear whether double reading was blinded. Discordant opinions were resolved via consensus or arbitration in 84.8% cases; 15.2% were double readings without consensus. The procedure of arbitration is not described.

Gromet (2008) [USA] analyzed the performance of double reading vs. the performance of the first reader in a double-reading program. Cases classified as negative by the first reader and positive by the second reader were resolved by the third reader who made the final decision. The first reader's recall rate was 10.2%, and the recall rate based on the final decision was 11.9%. Cancer detection rate was 4.12 per 1000 screens based on the first reader's opinion and 4.46 per 1000 screens based on the final decision. The estimated benefit of double reading was 38 additional cancers detected at a cost of 2,008 additional recalls and 140 additional biopsies; PPV decreased from 4.1% to 3.7% and the cancer detection rate increased by 0.34 per 1000; sensitivity increased from 81.4% to 88.0%. The author also analyzed the benefits of single reading with CAD and concluded: "With manpower and cost constraints limiting the use of double reading in the United States, CAD appears to be an effective alternative that provides similar, and potentially greater, benefits."

Posso et al. (2016) [Spain] found that blinded double reading was associated with a greater rate of false positive results (4.5% vs. 4.2%; P=0.001) compared to single reading. Cancer detection rates were similar with the two reading approaches (4.6 vs. 4.2 per 1000; P=0.283). In this screening program, discordant results were resolved by consensus. When the two readers could not reach a consensus, arbitration was undertaken by a third senior radiologist.

Double reading of potential recalls only

Two intervention studies were conducted to evaluate this strategy. An intervention study by **Mullen et al. (2017)** demonstrated that targeted utilization of double reading, specifically double reading of only potential recalls (consensus recall) with a third reader resolving disagreements, was efficient and not time consuming. Consensus recall was associated with significantly increased positive predictive value of recall, and this effect was seen with both full-field digital mammography (FFDM) and digital breast tomosynthesis (DBT). A significant decrease in recall rate after the introduction of consensus recall was seen only with FFDM. The authors explain the smaller effect with DBT by already reduced recall rates associated with DBT and more opportunity for improvement with FFDM. Also, the targeted double reading intervention was preceded by another intervention (so called "awareness" intervention that included performance feedback and review of personal recalls). The awareness intervention had a greater effect on recall rates than consensus recall, and the authors suggested that "there was marginal remaining opportunity after the awareness phase, therefore decreasing the additional opportunity available for improvement with consensus recall". Double reading of potential recalls took, on average, 2.3 minutes per case, including consultation with the third reader in case of disagreement. This study was conducted in an academic institution with breast imaging specialists. The authors acknowledge that the results may not be applicable outside these settings.

The poster presentation by **Rochman et al**. **(no date)** describes a Practice Quality Improvement Project (PQI) aimed at improving performance at the University of Virginia Department of Radiology and Medical Imaging. Screening mammograms at the institution were read by a group of four radiologists. The group identified screening mammography recall rates as an area for improvement and established an initial target of 10-12% for these rates. Recall rates and cancer detection rates that were collected from the mammography reporting system Megaview, were anonymized and distributed at the monthly faculty meeting. Root cause analysis was conducted to identify factors associated with increased rates of recall by individual radiologists. Identified potential causes included fear of missing a cancer, years of experience and recent implementation of tomosynthesis. The intervention consisted of double reading of all screening abnormal calls. All BI-RADS 0 examinations were independently reviewed by a different radiologist. If there was a disagreement, the case was discussed. The primary reader was responsible for the final impression and BI-RADS assessment category in each case. If the mammogram was categorized as BI-RADS 1 or 2, the names of both radiologists were issued on the report with the primary reader as the "reader" and the reviewer as an "agreer". Recall rates and cancer detection rates were collected monthly. Cancer detection rates were collected >30 days after the designated period to allow time for diagnosis. The combined screening recall rate for the four radiologists was 17.34% (range 15.47-20.80%) before the intervention. This rate decreased to 10.97% (range 10.37%-11.35%) during the first study cycle and maintained at 11.19% and 11.86% during the subsequent cycles. Cancer detection rates per 1000 were 6.5 before the intervention, 4.3 (first study cycle), 5.2 and 6.1 (subsequent cycles). The conclusion of this study: "Screening recall rates were reduced and maintained to the desirable level by implementation of this PQI initiative. Although recall rates were reduced, we did not experience a negative impact on the cancer detection rates for the group."

Study with inadequate characterization of reading/recall strategy

Salas et al. 2011 [Spain] found a higher false positive rate associated with double reading compared to single reading (OR=1.36; 95% CI: 1.23, 1.51). Double reading approach (e.g., blinded or non-blinded, consensus/arbitration or unilateral recall) is not described.

Overall Summary: double reading vs. single reading

The meta-analysis by Taylor and Potts (2008) demonstrated that, compared to single reading, double reading with unilateral recall or with mixed recall practice was associated with increased recall rates, while double reading with consensus or arbitration was associated with reduced recall rates. The authors acknowledge significant heterogeneity regarding the effect of double reading on recall rates. The original studies identified for this report show that, compared to single reading, double reading with either consensus/arbitration or unilateral recall increased the rate of recall/false positive results. Two intervention studies conducted in academic institutions demonstrated that targeted double reading of only potential recalls reduced recall rates without negatively affecting cancer detection rates.

The meta-analysis by Taylor and Potts (2008) found that cancer detection rates were increased at double reading compared to single reading regardless of recall strategy. The original publications demonstrated that double reading was associated with increased or unchanged cancer detection compared to single reading.

As discussed in several publications (**Ciatto et al. 2005**; **Posso et al. 2016**), the performance of double-reading vs. single reading may depend on the experience of screen readers. For example, **Ciatto et al. (2005)** suggested that benefits from double reading might be larger if less experienced screen readers were involved.

Approaches to double reading: blinded vs. non-blinded double reading
With the introduction of full field digital mammography (FFDM) it became technically possible to perform blinded double reading instead of non-blinded double reading (Klompenhouwer, 2015a).

Summary of Review Articles
In their narrative review, Le et al. (2016) concluded that blinded double reading was associated with a higher rate of false positives as compared to non-blinded double reading. Blind reading improves cancer detection.

Summary of Original Studies
Klompenhouwer, 2015a [the Netherlands] observed significantly higher recall, false positive and cancer detection rates at blinded vs. non-blinded double reading. Women with discrepant readings between the two radiologists, at blinded and non-blinded double reading, were always recalled for further analysis.

Overall Summary: blinded vs. non-blinded double reading
The limited information identified for this report suggests that blinded double reading increases recall, false positive and cancer detection rates compared to non-blinded double reading.

Approaches to double reading: consensus/arbitration of discordant opinions vs. unilateral recall

Summary of Review Articles
In their systematic review, Hackney et al. (2017) concluded:

1) "Overall, studies reported that compared to highest reader recall (non-arbitration), arbitration resulted in significant reductions in recall rates, with relative decreases in the range of 17.8%...to 40.9%."
2) "There is disparity between the studies regarding the effect of arbitration on cancer detection rates."

Summary of Original Studies
Caumo et al. 2011b [Italy] found that, compared to unilateral recall in a double reading program, arbitration of discordant opinions was associated with 2.8% absolute reduction in recall rate and 40.9% relative reduction. An estimated absolute reduction in cancer detection rate by arbitration of discordant opinions would be 0.13 per 1000 (relative reduction 2%). The authors estimated that the cost of arbitration was 74 euros, and the cost of 216 spared assessment procedures was between 14,558.4 and 23.346 euros. The authors concluded: "Arbitration is a cost-effective procedure that could be employed as a first measure to counterbalance excess recall rate observed in a double-reading scenario." From the description of this study, it is unclear whether the second reader was aware of the first reader's opinion, and whether the arbitrator was aware of the reason for discrepancy.

Klompenhouwer et al, 2015b [the Netherlands] found that, both in blinded and in non-blinded double reading, arbitration was associated with significantly lower recall rates, significantly higher positive predictive values, without a significant change in the cancer detection rates. Arbitration resulted in a reduction in programme sensitivity, and this effect was statistically significant at blinded double reading. In this study, the arbitrator was blinded to the screening outcome.

When BI-RADS 0 recalls only were arbitrated **(Klompenhouwer, 2015c**) recall rates at blinded and non-blinded double reading were decreased without a decrease in cancer detection rate and sensitivity. The positive predictive value was increased by arbitration of BI-RADS 0 recalls at blinded double reading.

Posso et al. (2016) found that blinded double reading with consensus or arbitration was associated with lower recall rates compared to blinded double reading with unilateral recall (4.5% vs. 6.0%; P<0.001). Cancer detection rates were not different with the two approaches. In this study, arbitration by a third senior radiologist was undertaken if the two readers could not reach consensus.

Shaw et al. (2009) [Ireland] analyzed recall rates under three different scenarios in a screening program with independent double reading: highest reader recall when a woman is recalled if her findings are deemed abnormal by either reader [the term "highest reader recall" appears to be equivalent to the term "unilateral recall"]; unanimous recall, when none of the patients with discordant findings was referred for further assessment, and consensus review of discordant findings. Consensus review was associated with a slightly lower recall rate compared to the highest reader recall (4.41% vs. 4.97%) and a slightly higher recall rate compared to unanimous recall (4.41% vs. 3.94%). Cancer detection rate was slightly lower with consensus review compared to the highest reader recall (7.47 vs. 7.53 per 1000). In this study, discordant opinions were resolved by a consensus panel that met twice a week. The panel consisted of 3 to 5 radiologists and usually included one or both original researchers. A woman was recalled if any member of the panel recommended referral. Although the method of resolution of discordant opinions in this study is described as a consensus review, it resembles arbitration. It should be noted that, in their systematic review, Hackney et al (2017) acknowledged difficulties in differentiating between the effects of arbitration and consensus on recall rates: the original studies either did not provide a clear definition of consensus and arbitration, or the two processes were integrated in the decision making. Regarding consensus panels, Hackney et al (2017) noted that "the dynamics within the consensus team can be a significant factor affecting the final decision." For example, "one reader is the dominant and opinions are not equally weighted" or "individuals may change their judgment to what they 'believe others want to hear'".

Overall Summary: consensus/arbitration of discordant opinions vs. unilateral recall
Resolution of discordant opinions via consensus or arbitration appears to be associated with reduced recall rates compared to unilateral recall at double reading. Cancer detection rates were either similar with the two approaches to double reading, or slightly lower with consensus/arbitration.

Approaches to double reading: the order in which two readers examine a batch of mammograms
The aim of a multi-center, double-blind, cluster randomized clinical trial conducted in the UK (**Taylor-Phillips et al. 2016**) was to determine whether a vigilance decrement (reduced detection rate with time on task) exists in breast cancer screening and whether changing the order in which two readers examine a batch of mammograms can increase the cancer detection rate (assuming that the two readers experience peak vigilance at different points within the reading batch). The two readers examined each batch of mammograms in the same order (control group) or in the opposite order to one another (intervention group). The following conclusion was reached by the researches: "The intervention did not influence cancer detection rate, recall rate, or rate of disagreement between readers. There was no

pattern of decreasing cancer detection rate with time on task as predicted by previous research on vigilance decrements as a psychological phenomenon. Instead there was a gradual decrease in recall rate, with an increase in PPV and a decrease in false-positive recall of women with time on task."

Audit/Performance feedback

<u>*Summary of Review Articles*</u>
In their narrative review, **Soh et al. (2012)** focused on limitations of two methods for monitoring accuracy of interpretation and providing feedback to screen readers: clinical audit and standardized screening test sets. Although clinical audit is used with good effect to assess screen readers' performance, it may take a long time (up to 2 years) for falling performance to be identified by audit and another 2 years to demonstrate improvement in performance after introduction of quality improvement plans. This may take even longer for low volume breast screening programs. Standardized test sets, such as the Personal Performance in Mammographic Screening (PERFORMS) test implemented by the NHSBSP (UK), and BREAST (Breastscreen Reader Assessment STrategy) introduced in Australia, have several advantages. These include ease of application, immediate feedback, and quicker assessment of quality improvement measures. However, the test results require validation against real clinical reading performance. The authors identified four key factors that impact the external validity of screening test sets: "the nature and extent of scrutiny of one's action, the artificiality of the environment, the oversimplification of responses, and prevalence of abnormality".

<u>*Summary of Original Studies*</u>
Of the six articles identified for this report, three are not informative regarding the potential influence of performance feedback on recall rates[6]. The three informative articles describe two intervention studies.

Mullen et al. 2017 [USA] evaluated the effectiveness of a so-called awareness intervention that included two phases. In phase 1, each radiologist compared his/her individual performance to that of the group. The group discussed perceptions of recall/performance, such as most frequent reasons for recall and individual fears prompting recall. A goal was set to reduce the group's and each radiologist's recall rate to 5%, while monitoring cancer detection rate and PPV. In phase 2, each radiologist weekly reviewed the imaging and reports of his/her recalls, and then the imaging and reports from the subsequent diagnostic evaluation/biopsy for each recalled patient. This was a long-term intervention; it continued for seven months from February 3 to September 3, 2015. The Awareness intervention was associated with significantly decreased recall rates, and the decrease was seen with both full-field digital mammography (FFDM) and with digital breast tomosynthesis (DBT). The intervention had no significant effect on cancer detection rates, and the positive predictive value was significantly increased with DBT. The authors

[6] **Geertse, 2015** reports on recall rates, cancer detection rates and positive predictive values of recall over four periods corresponding to four series of audit performed by the Dutch Reference Center for Screening in seventeen Dutch mammogram reading units. The authors also describe how the audit program works. Because changes occurred during the study period in mammography technology, techniques and mammogram reading practices, it is unclear if the observed trends in performance indicators were associated with the audit/performance feedback. **Hofvind, 2016** describe results of a web-based survey on audit feedback conducted in 17 screening programs in member countries of the International Cancer Screening Network (ICSN). There is no analysis of possible influence of audit feedback on program performance indicators. Likewise, **Lester and Dall (2003)** report on audit results (performance of five radiologists) but there is no analysis of possible influence of the audit on the performance of these radiologists. More details on these studies can be found in the Appendix.

concluded that simple interventions, such as personal review of recalls had the potential to decrease recall rates. However, the authors acknowledged that, because their study was conducted at an academic institution with breast imaging specialists, the results might not be applicable outside of an academic subspecialty settings. Another limitation discussed by the authors was the small sample size.

Carney et al. (2011, 2012) [USA] developed and implemented an interactive web-based intervention that included three components (modules): 1) Peer comparison audit data on performance indicators; this module was also aimed at explaining audit statistics and how they were derived; 2) Addressing radiologists' misconception about women's' risk of breast cancer; 3) Addressing radiologists' misconceptions regarding malpractice related to breast imaging. This was a short-term intervention: it took on average about one hour to complete all three modules. Most radiologists found the program moderately to very helpful, and the percentage of radiologists who reported that the risk of medical malpractice influenced their recall rates dropped considerably (from ≈36 pre-intervention to ≈18% post-intervention). However, the intervention had no effect on recall rates. The authors believed that a single intervention might not be adequate to address excessive recall rates, and that more complex approaches might be needed to change patterns of radiologists' practice.

Overall Summary

In summary, according to the review article by **Soh et al. (2012),** audit may help improve radiologists' performance; however, it may take up to several years for the improvements to be seen. Standardized test sets provide immediate feedback and quicker assessment of quality improvement measures but their results require validation against real clinical reading performance. Limited information has been identified on the effectiveness of interventions that included feedback and educational components. Based on this limited information, it is not possible to determine factors associated with effectiveness of such interventions. It is possible that longer-term interventions are more successful in reducing excessive recall rates.

Comparison with Prior Mammograms

Summary of Review Articles

No relevant review articles have been identified.

Summary of Original Studies

Klompenhouwer et al. (2014) [the Netherlands] showed that, during the transition from screen-film (SFM) to full-field digital mammography (FFDM), there was a significant increase in the proportion of women who had been recalled twice for the same mammographic lesion. Breast cancer was significantly less often diagnosed in these women than at SFM. Blinded review demonstrated that, availability of an older hard copy SFM examination in addition to the most recent digitalized SFM examination at the first round of FFDM screening would have reduced the number of women repeatedly recalled for the same lesion by almost 40%. This finding was important because the rate of re-attendance at screening was significantly lower for women who had had a repeated false positive recall, especially if both recalls were for the same mammographic lesion.

Hayward et al. (2016) [USA] demonstrated that, when two or more prior examinations were used for comparison with the current examination, recall rate significantly decreased while the PPV and CDR increased relative to comparison with a single prior examination. During the period of this study, only digital mammography was used; however, prior examinations also included screen film. It is not clear whether prior screen films were digitalized for display.

When converting an analogue to a digital image, there is a loss of image information due to pixellation (**Taylor-Phillips et al. 2012**). The aim of **Taylor-Phillips et al. (2012**) [UK] was to study the effect of the presentation medium of the prior mammograms on the performance using digital mammography. Specifically, the performance was examined with film prior mammograms, digitalized prior mammograms or without prior mammograms. This study was performed in a test set of mammograms, and the quantitative results were projected to a real-life scenario. The number of false positive cases was significantly higher without prior mammograms than with prior mammograms, and there was no significant difference between using film or digitalized format of prior mammograms. The 26% increase in false positives when prior mammograms were <u>not</u> used, relative to when they were used in either format, would correspond to an increase in recall rate at the study hospital from 4.3% to 5.5% with no associated increase in cancer detection. The estimated cost of this increase was higher than the cost of displaying prior mammograms. The authors acknowledge that their findings may not translate into equivalence of performance using film and digitized prior mammograms in a real-life screening situation: the participants were aware that they were reading difficult cases, which could result in greater vigilance. Also, reading an enriched test set with a greater proportion of abnormal cases than in screening practice may have led to an underestimate of recalls in screening practice.

Yankaskas et al. (2011) [USA] observed higher recall rates and higher cancer detection rates in women for whom prior mammograms were available and used for comparison with the current examination, relative to women with no prior mammograms. This comparison may not be meaningful because the proportion of prevalent examinations is higher among examinations with no prior mammograms, and higher recall/cancer detection rates should be expected. The authors also demonstrated that, if there was a change on (from-?) the prior mammogram, recall rates and cancer detection rates were higher compared to no changes on prior mammograms. The meaning of “change” is unclear because, in some cases, the authors refer to “change <u>on</u> comparison mammograms” or “change(s) <u>in</u> the comparison image”, in other cases they refer to “change <u>from</u> the comparison mammogram”.

<u>*Overall Summary*</u>
In summary, studies identified for this report demonstrate that comparison with prior mammograms is associated with reduced recall rates with no negative impact on cancer detection. When two or more prior examinations are used for comparison with the current examination, recall rates may be lower relative to comparison with a single prior examination. Reduction in recall/false positive rates may be of similar magnitude regardless of whether prior mammograms are digitalized or displayed as films.

Number of mammographic views
This factor was discussed by **Le et al. (2016)** in their narrative review. Acquisition of two views (craniocaudal and mediolateral oblique) significantly decreases the rate of false positives as compared to

a single view. Because screening mammography in the US and Canada has been conducted using two views since the 1980's, it is unlikely that this practice contributed to the higher false positive rates in North America (Le et al. 2016).

In consultation with the Partnership, publications reporting on original research relevant to this factor were not considered in this report.

Mammographic compression
A single study examining this factor was identified. **Holland et al. (2016)** [The Netherlands] divided 113,464 screening examinations into five groups defined by compression pressure applied during the acquisition of the mammogram. These groups were ≤7.68 kPa; >7.68 to ≤9.18 kPa; >9.18 to ≤10.71 kPa; >10.71 to≤12.81 kPa and >12.81 kPa. Significant differences across the five groups were seen for the positive predictive values and cancer detection rates with the moderate pressure groups having higher rates compared to the first and last groups. Similar, although not statistically significant trend was seen for the false positives. No trend was seen for recall rates. The authors concluded that "too low or too high compression may reduce screening program performance". The authors suggest that European guidelines be "more specific in their recommendations, going from a descriptive recommendation ('firm but tolerable') to a quantitative recommendation in kPa".

Batch reading of mammograms

<u>*Summary of Review Articles*</u>
No relevant review articles have been identified.

<u>*Summary of Original Studies*</u>
Burnside et al. (2005) [USA] compared recall and cancer detection rates before and after introduction of batch reading. In batch reading, dedicated uninterrupted distraction-free time was provided for interpreting screening mammograms. Telephones in the reading room were changed to lines for outgoing calls only. Before the introduction of batch reading, the so-called non-batch reading offline (i.e., after the patient left the premises) was practiced. At non-batch reading offline, no consistent dedicated time was available for interpretation of screening mammograms. Screening examinations were interpreted between other activities; the interpretation was routinely interrupted by telephone calls, diagnostic imaging and other activities. The study demonstrated that introduction of batch reading resulted in a significant reduction in recall rates without affecting cancer detection rates. Subset analyses conducted to control for possible confounding effect of changes in technology during the study period (switch to digital mammography and introduction of CAD) demonstrated that, most likely, these changes did not play a role in decreasing the recall rates. Batch reading resulted in a decrease in recall rates both with analog and digital mammography. However, because fewer mammograms were acquired with digital mammography, the authors concluded that the effect of batch reading on digital mammography needed further clarification.

The study by **Ghate et al. (2005)** [USA] is different from **Burnside et al. (2005)** in that non-batch reading of mammogram (referred to as immediate reading) was performed while the patient was still waiting for the results. The results were communicated to the patient at the time of the visit, and any necessary additional imaging was also performed during this visit. In batch reading, mammograms were interpreted in a batch reading session after the patient left. It is not clear whether the batch reading sessions were distraction-free. Also, the batch and non-batch reading were used concurrently with the radiologists rotated evenly between assignments for immediate and batch reading of the mammograms. Delayed batch reading of mammograms was associated with lower recall rates than immediate reading. Cancer detection rates were similar with these two reading approaches.

Overall Summary
In summary, two studies conducted in the USA demonstrated that recall rates associated with batch reading were lower than recall rates associated with immediate (online) or offline non-batch reading. Cancer detection rates were unaffected by this practice. One of these studies stressed the importance of uninterrupted distraction free environment during batch reading sessions.

Radiologist characteristics

Training, Education, and Experience

Summary of Review Articles
In most countries, screening mammograms are interpreted by radiologists. However, radiographers in the United Kingdom (UK) and breast physicians in Australia also act as screen readers (**Mohd Norsuddin et al. 2015)**. van den **Biggelaar (2008)** performed a systematic review of the literature focused on the performance of radiographers (technologists and physician assistants) compared with radiologists in the interpretation of mammograms. Six studies published in the 1980's to 1990's met the inclusion criteria. Compared to radiologists, radiographers had higher false positive rates with similar sensitivity in the detection of malignancies in a screening setting. These results suggest that reading performance can improve with training.

Summary of Original Studies

Years of practice and fellowship training
Several studies suggest that reader's performance in screening mammography improves throughout their career. For example, **Alberdi et al. (2011)** [Spain] showed that the risk of overall false-positive results and false-positives leading to an invasive procedure decreased with increasing years of service in screening mammography. Analyses by **Barlow et al. (2004)** [USA] showed that recall rates significantly decreased with increasing years of mammography interpretation. When all statistically significant radiologist's characteristics were included in the model, increasing number of years in mammography practice was significantly associated with increasing specificity. (Recall rates were not modelled in this way). **Smith-Bindman et al. 2005** [USA] demonstrated a decline in false-positive rates and an improvement in specificity with increasing time since receipt of medical degree (which was likely used as a surrogate for duration of practice as a screen reader). **Tan et al. (2006)** [USA] also found that more recently trained radiologists had higher false-positive rates.

Elmore et al. (2009) [USA] found significantly lower recall and false-positive rates among radiologists with 10 to 19 years of experience in interpreting mammograms compared to those with less than 10 years of experience. However, there was no consistent increasing trend: recall and false-positive rates of radiologists with 20 or more years of experience were similar to those of radiologists with less than 10 years of experience.

Cornford et al. (2004) [UK] found no significant association between years of experience and any of the performance outcome measures, including recall rates and cancer detection rates. This study included only 37 screen readers, of whom 16 were radiographers. The authors acknowledged that their results were “likely to be affected by occupational group”. As well, it appears that the analyses were not adjusted for any patient’s or reader’s characteristics; this was likely due to the small sample size.

Carney et al. (2004) [USA] assessed radiologists’ reactions to uncertainty and found that more experienced interpreters had lower reactions to uncertainty than radiologists who were new to practice. The findings suggest that reactions to uncertainty lessened with more years of experience. Higher uncertainty scores were associated with increased recall rates, although not significantly.

The effect of the number of years of image interpretation on reader’s performance may be modified by fellowship training in breast imaging. **Miglioretti et al. (2009)** [USA] demonstrated that radiologists who received fellowship training in breast imaging did not have a learning curve; they reached the Agency for Healthcare Research and Quality (AHRQ) desirable performance goals for screening mammography during their first year of practice. Recall rates, false-positive rates or PPV1 did not change significantly with increasing years of experience. In contrast, recall and false-positive rates for radiologists without fellowship training were significantly higher than the AHRQ desirable goals during the first year of practice. Only radiologists with 19 or more years of experience had recall and false positive rates meeting the AHRQ desirable goals. Because the largest improvement in the interpretive performance occurred during the first 3 years of practice, **Miglioretti et al. (2009)** [USA] concluded that educational interventions, system-level support (e.g., double reading with consensus and arbitration) and feedback on radiologists’ interpretive performance, could be especially important during the first years of practice.

In contrast to the study by **Miglioretti et al. (2009)** which analyzed within-radiologist effects over time (by period of radiologists’ career), the study by **Elmore et al. (2009)** compared performance indicators of the two groups (radiologists with and without fellowship training) over their entire careers. These two articles had similar authorship, reported on data from the same seven registries contributing to the Breast Cancer Surveillance Consortium, and were published in the same issue of the Radiology journal. **Elmore et al. (2009)** found significantly <u>higher</u> recall and false-positive rates among fellowship-trained radiologists than in radiologists who had no fellowship training. Additionally, fellowship-training was found to be associated with higher PPV, cancer detection rates, sensitivity, and overall accuracy (**Elmore et al. 2009**). In the study by **Miglioretti et al. (2009)** and **Elmore et al. 2009**., fellowship-trained radiologists constituted only 7-8% of the sample analyzed.

Working full time vs. part time

In analyses by **Barlow et al. (2004)** [USA], recall rates were not significantly different between radiologists working full time and those working part time.

Percent of time spent/hours per week working in breast imaging

Barlow et al. (2004) [USA] did not see a consistent trend in recall rates with increasing percent of time spent working in breast imaging. Specifically, radiologists spending 20 to 39% of their time in breast imaging had significantly higher recall rates compared to radiologists who spent less than 20% of their time in breast imaging (reference group). However, compared to the same reference group, significantly lower recall rates were observed among those who spent 40% or more of their time in breast imaging.

Elmore et al. (2009) [USA] found no significant trend in recall rates with increasing hours/week working in breast imaging

Affiliation

Barlow et al. (2004) [USA] found that affiliation with an academic medical center had no effect on recall rates. In contrast, **Elmore et al. (2009)** [USA] observed significantly lower recall rates in radiologists affiliated with academic medical centers (adjunct or primary affiliation). Also, radiologists who had adjunct affiliations with an academic medical center had lower false positive rates.

Experience with tomosynthesis

DiPrete et al. (2018) [USA] measured radiologists' performance in a community practice where only digital mammography was available, before and after having experience with digital breast tomosynthesis (DBT). Both recall rates and cancer detection rates of digital mammography increased after radiologist's experience with DBT. [Note: this article addresses the effect of experience with DBT while continuing working with digital mammography, not differences in performance between the two technologies].

Other factors related to experience

Tan et al. 2006 [USA] found no significant effect of the type of practice (indirect patient care vs. direct patient care) or board certification in radiology on false positive rates.

Overall summary

Overall, although most studies show that increasing number of years in mammography is associated with decreasing recall/false positive rates, there are studies that do not find such an association or do not show a consistent trend. **Miglioretti et al. (2009)** noted a common limitation of studies on length of service: they compare groups of radiologists who had interpreted mammograms for different periods of time rather than analyzing changes within individual radiologists over time. Comparisons between the groups defined by length of service can be confounded by changes in medical education and practices over time and by differences between radiologists who had decided to stay in mammography for many years and those who had recently entered the field. One study suggests that the effect of length of service on radiologist's performance may be modified by fellowship training in breast imaging. Specifically, fellowship-trained radiologists did not have the learning curve characteristic of radiologists who were not fellowship-trained. Another study of the same population demonstrated higher recall rates, false-positive rates, sensitivity, and overall accuracy over the entire career of radiologists with fellowship training compared to those without. Two studies provided inconsistent findings regarding the effect on recall rates of affiliation with academic medical centers. Previous experience with digital breast tomosynthesis while continuing work with digital mammography, was associated with increased recall rates in one study.

Based on the results of a single study, there is no evidence for the effect of other factors related to radiologists' experience (percent of time spent/hours per week working in breast imaging, working full time vs. part time, indirect patient care vs. direct patient care) on recall rates.

Radiologists' demographics

Age

In analyses adjusted for patients' characteristics (**Barlow et al. 2004** [USA]), a significant inverse association was identified between recall rates and the age of radiologists. However, no significant relationship between the age of radiologists and sensitivity or specificity were observed after including all statistically significant radiologists' factors in the mixed-effect models.

In analyses by **Smith-Bindman et al. (2005)** [USA], false-positive rates declined (i.e., specificity improved) with increasing age of radiologists. As the rate of false positives also declined with increasing time since receipt of medical degree (a surrogate for duration of service), it may not be easy to disentangle these two effects. In an unadjusted analysis by **Tan et al. (2006)** [USA], rates of false-positive results significantly decreased with increasing radiologists' age. Because a similar significant decreasing trend in false-positive rates was observed with increasing number of years since graduation (surrogate for service duration), the observed effect may be related to gaining experience in mammogram reading rather than to radiologists' age.

Gender

Elmore et al. (2009) found a significantly higher recall and false positive rate and significantly lower positive predictive value among female radiologists. There were no significant gender differences in sensitivity. These analyses were adjusted for patients' characteristics and radiologists' characteristics. **Tan et al. (2006)** also found higher false-positive rates among female radiologists. This analysis was also adjusted for patients' and radiologists' characteristics. **Barlow et al. (2004)** found no association between radiologists' gender and recall rates or other performance measures. Analyses of gender effects were adjusted for patients' characteristics but not for radiologists' characteristics (e.g., experience). Only statistically significant radiologist factors were then tested together using mixed-effects models, and because gender was not a significant factor, it was most likely not included in these models. The authors concluded that radiologists' gender was not associated with performance. This study includes fewer radiologists than **Elmore et al. (2009)** or Tan et al. **(2006)**. **Carney et al. (2004)** provided some "indirect" evidence that female radiologists may have lower recall rates. These researchers found that male radiologists reported more intense reactions to uncertainty and had a higher mean combined uncertainty score than female radiologists. Higher uncertainty scores were associated with increased recall rates, although the association was not statistically significant.

Overall Summary

Overall, several studies demonstrate decreasing recall/false positive rates with increasing radiologists age. However, this decline may be related to experience rather than age. Few studies have been identified that report on the potential association between radiologists' gender and performance. Two publications identified for this review suggest that female radiologists have higher recall/false positive rates. These conclusions come from analyses of relatively large populations and adjusted for other radiologists' characteristics such as experience. One study of a smaller size found no statistically significant differences between male and female radiologists in analyses not adjusted for other radiologists' characteristics. The evidence from one study that female radiologists may perform better than males in terms of recall rates is indirect and is not statistically significant.

Litigation concern

Summary of Review Articles

In their narrative review**, Le et al (2016)** reported that, in the USA, perceived risk of litigation associated with medical malpractice had been identified as a possible contributor to the relatively high recall rates. These concerns were also present among Canadian radiologists: 72% of surveyed radiology residents reported a strong concern of malpractice risk specific to mammography as compared with other imaging examinations. **Le et al (2016)** concluded in their review that "litigation risk should not be discounted as a potential contributing factor to the heightened recall rates observed in North America".

Summary of Original Studies

Whang et al. (2013) studied the most frequent causes of malpractice suits using data on 8401 American radiologists practicing in 47 states. The most common cause was error in diagnosis, and the most frequently missed diagnosis was breast cancer. **Barlow et al. (2004)** analyzed sensitivity and specificity in relation to variables characterizing malpractice experience and concerns. None of these variables were significantly associated with radiologists' performance measures. **Carney et al. (2004)** found that radiologists with any prior medico-legal experience had slightly (not significantly) higher uncertainty scores. Although higher uncertainty scores were associated with increased recall rates, the association was not statistically significant. **Elmore et al. (2005)** conducted a mail survey among radiologists interpreting mammograms to assess the relationship between radiologists' perception of and experience with medical malpractice and their recall rates. This study demonstrated that U.S. radiologists were extremely concerned about medical malpractice. They reported that this concern affected their recall rates and recommendations for biopsy. However, variables characterizing medical malpractice experience and concerns were not associated with recall or false-positive rates. **Carney et al. (2011, 2012)**[7] conducted an intervention that included, as one of the components, addressing radiologists' misconceptions regarding malpractice related to breast imaging. The percentage of radiologists who reported that the risk of medical malpractice influenced their recall rates dropped from 36.3% pre-intervention to 17.8% post-intervention. However, no associated decrease in recall rates was observed.

Overall Summary

Overall, although radiologists reported they were concerned about medical malpractice, variables characterizing medical malpractice experience and concerns were not associated with recall or false-positive rates.

[7] Details on this study can be found in the data abstraction table Quality Assurance Practices (section Audit/Performance Feedback").

Summary of Main Findings

Based on the analysis of available evidence, the factors under consideration may be classified into four groups.

Factors that may decrease recall rates without compromising cancer detection

- Implementation of digital breast tomosynthesis (DBT) in screening practice
- Targeted double reading of only potential recalls
- Consensus or arbitration vs. unilateral recall at double reading
- Comparison with two or more prior mammograms
- Batch reading of mammograms
- Fellowship training in breast imaging

Factors that may merit further consideration in designing breast cancer screening programs

- Synthesized mammography. This factor was not analyzed in depth within the framework of this project. However, initial screen of literature suggests that, used as an adjunct to DBT, this technology may preserve the performance benefits provided by DBT and at the same time reduces the dose of radiation.
- Interventions that include performance feedback and educational components are potentially effective in decreasing recall rates while maintaining cancer detection rates. Factors that determine their effectiveness need to be identified.
- Reading volume: Although overall evidence is inconclusive, a Canadian study of good quality demonstrates gains in interpretive accuracy with increasing reading volume; the gain is greater in the range of reading volumes up to about 3000 mammograms per year.
- Mammographic compression: evidence from a single study shows that false positive rates may be lower and cancer detection rates are significantly higher at moderate compression pressure compared to low or high pressure.

Non-modifiable factors that may influence recall rates

- Recall rates may decrease with increasing years of experience interpreting mammograms
- Female radiologists tend to have higher recall rates than male radiologists.

Factors with inconsistent or insufficient evidence on their effect on recall rates

- Introduction of digital mammography
- Computer assisted detection systems (CAD). There are different ways in which CAD is used (e.g., as a second reader, as an arbitrator of discordant opinions), and it can be used as an adjunct to different technologies. The effect of CAD on performance may differ depending on the way it is used as well as on the experience of screen reader who is using it. Little research is available to address these aspects of CAD use. It should also be noted that manufacturers of CAD systems work on improvements of CAD algorithms to increase specificity by reducing false prompts.

References

Agbaje, O. F., Astley, S. M., Gillan, M. G. C., Boggis, C. R. M., Wilson, M., Barr, N. B., et al. (2006) Mammography reading with Computer-Aided Detection (CAD): Single view vs two views. *Vol. 4046 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 125-130).

Alberdi, R. Z., Llanes, A. B., Ortega, R. A., Exposito, R. R., Collado, J. M., Verdes, T. Q., et al. (2011). Effect of radiologist experience on the risk of false-positive results in breast cancer screening programs. *Eur Radiol, 21*(10), 2083-2090.

Ambinder, E. B., Harvey, S. C., Panigrahi, B., Li, X., & Woods, R. W. (2018). Synthesized Mammography: The New Standard of Care When Screening for Breast Cancer with Digital Breast Tomosynthesis? *Acad Radiol*.

Aujero, M. P., Gavenonis, S. C., Benjamin, R., Zhang, Z., & Holt, J. S. (2017). Clinical Performance of Synthesized Two-dimensional Mammography Combined with Tomosynthesis in a Large Screening Population. *Radiology, 283*(1), 70-76.

Bargallo, X., Santamaria, G., Del Amo, M., Arguis, P., Rios, J., Grau, J., et al. (2014). Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. *Eur J Radiol, 83*(11), 2019-2023.

Barlow, W. E., Chi, C., Carney, P. A., Taplin, S. H., D'Orsi, C., Cutter, G., et al. (2004). Accuracy of screening mammography interpretation by characteristics of radiologists. *J Natl Cancer Inst, 96*(24), 1840-1850.

Bennett, R. L., Sellars, S. J., Blanks, R. G., & Moss, S. M. (2012). An observational study to evaluate the performance of units using two radiographers to read screening mammograms. *Clin Radiol, 67*(2), 114-121.

Bernardi, D., Macaskill, P., Pellegrini, M., Valentini, M., Fanto, C., Ostillio, L., et al. (2016). Breast cancer screening with tomosynthesis (3D mammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study. *Lancet Oncol, 17*(8), 1105-1113.

BreastScreen Australia. National Quality Improvement Plan 2018-2020. Approved by the National Quality Management Committee on 2 March 2018. http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/EFBA481DF32AA093CA257CEC00220FC4/$File/NQIP%2018-20%20-%20Final.pdf

Buist, D. S., Anderson, M. L., Haneuse, S. J., Sickles, E. A., Smith, R. A., Carney, P. A., et al. (2011). Influence of annual interpretive volume on screening mammography performance in the United States. *Radiology, 259*(1), 72-84.

Burnside, E. S., Park, J. M., Fine, J. P., & Sisney, G. A. (2005). The use of batch reading to improve the performance of screening mammography. [Research Support, Non-U.S. Gov't]. *AJR. American Journal of Roentgenology, 185*(3), 790-796.

Campari, C., Giorgi Rossi, P., Mori, C. A., Ravaioli, S., Nitrosi, A., Vacondio, R., et al. (2016). Impact of the Introduction of Digital Mammography in an Organized Screening Program on the Recall and Detection Rate. *J Digit Imaging, 29*(2), 235-242.

Carney, P. A., Abraham, L., Cook, A., Feig, S. A., Sickles, E. A., Miglioretti, D. L., et al. (2012). Impact of an educational intervention designed to reduce unnecessary recall during screening mammography. *Acad Radiol, 19*(9), 1114-1120.

Carney, P. A., Elmore, J. G., Abraham, L. A., Gerrity, M. S., Hendrick, R. E., Taplin, S. H., et al. (2004). Radiologist uncertainty and the interpretation of screening. *Med Decis Making, 24*(3), 255-264.

Carney, P. A., Geller, B. M., Sickles, E. A., Miglioretti, D. L., Aiello Bowles, E. J., Abraham, L., et al. (2011). Feasibility and satisfaction with a tailored web-based audit intervention for recalibrating radiologists' thresholds for conducting additional work-up. *Academic Radiology, 18*(3), 369-376.

Caumo, F., Brunelli, S., Zorzi, M., Baglio, I., Ciatto, S., & Montemezzi, S. (2011a). Benefits of double reading of screening mammograms: retrospective study on a consecutive series. *Radiol Med, 116*(4), 575-583.

Caumo, F., Brunelli, S., Tosi, E., Teggi, S., Bovo, C., Bonavina, G., et al. (2011b). On the role of arbitration of discordant double readings of screening mammography: experience from two Italian programmes. *Radiol Med, 116*(1), 84-91.

Caumo, F., Zorzi, M., Brunelli, S., Romanucci, G., Rella, R., Cugola, L., et al. (2018). Digital Breast Tomosynthesis with Synthesized Two-Dimensional Images versus Full-Field Digital Mammography for Population Screening: Outcomes from the Verona Screening Program. *Radiology, 287*(1), 37-46.

Chiarelli, A. M., Edwards, S. A., Prummel, M. V., Muradali, D., Majpruz, V., Done, S. J., et al. (2013). Digital compared with screen-film mammography: performance measures in concurrent cohorts within an organized breast screening program. *Radiology, 268*(3), 684-693.

Ciatto, S., Ambrogetti, D., Bonardi, R., Catarzi, S., Risso, G., Rosselli Del Turco, M., et al. (2005). Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening programme. *J Med Screen, 12*(2), 103-106.

Coldman, A. J., Major, D., Doyle, G. P., D'Yachkova, Y., Phillips, N., Onysko, J., et al. (2006). Organized breast screening programs in Canada: effect of radiologist reading volumes on outcomes. *Radiology, 238*(3), 809-815.

Comas, M., Arrospide, A., Mar, J., Sala, M., Vilaprinyo, E., Hernandez, C., et al. (2014). Budget impact analysis of switching to digital mammography in a population-based breast cancer screening program: a discrete event simulation model. *PLoS One, 9*(5), e97459.

Cornford, E., Reed, J., Murphy, A., Bennett, R., & Evans, A. (2011). Optimal screening mammography reading volumes; evidence from real life in the East Midlands region of the NHS Breast Screening Programme. *Clin Radiol, 66*(2), 103-107.

Dabbous, F., Dolecek, T. A., Friedewald, S. M., Tossas-Milligan, K. Y., Macarol, T., Summerfelt, W. T., et al. (2017). Performance characteristics of digital vs film screen mammography in community practice. *Breast J*.

de Munck, L., de Bock, G. H., Otter, R., Reiding, D., Broeders, M. J., Willemse, P. H., et al. (2016). Digital vs screen-film mammography in population-based breast cancer screening: performance indicators and tumour characteristics of screen-detected and interval cancers. *Br J Cancer, 115*(5), 517-524.

DiPrete, O., Lourenco, A. P., Baird, G. L., & Mainiero, M. B. (2018). Screening Digital Mammography Recall Rate: Does It Change with Digital Breast Tomosynthesis Experience? *Radiology, 286*(3), 838-844.

Duncan, K. A., & Scott, N. W. (2011). Is film-reading performance related to the number of films read? The Scottish experience. *Clin Radiol, 66*(2), 99-102.

Durand, M. A. (2018). Synthesized Mammography: Clinical Evidence, Appearance, and Implementation. *Diagnostics (Basel)*. 8(2).

Elmore, J. G., Jackson, S. L., Abraham, L., Miglioretti, D. L., Carney, P. A., Geller, B. M., et al. (2009). Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology, 253*(3), 641-651.

Elmore, J. G., Taplin, S. H., Barlow, W. E., Cutter, G. R., D'Orsi, C. J., Hendrick, R. E., et al. (2005). Does litigation influence medical practice? The influence of community radiologists' medical malpractice perceptions and experience on screening mammography. *Radiology, 236*(1), 37-46.

Feeley, L., Kiernan, D., Mooney, T., Flanagan, F., Hargaden, G., Kell, M., et al. (2011). Digital mammography in a screening programme and its implications for pathology: a comparative study. *J Clin Pathol, 64*(3), 215-219.

Fenton, J. J., Abraham, L., Taplin, S. H., Geller, B. M., Carney, P. A., D'Orsi, C., et al. (2011). Effectiveness of computer-aided detection in community mammography practice. *J Natl Cancer Inst, 103*(15), 1152-1161.

Friedewald, S. M., Rafferty, E. A., Rose, S. L., Durand, M. A., Plecha, D. M., Greenberg, J. S., et al. (2014). Breast cancer screening using tomosynthesis in combination with digital mammography. *JAMA, 311*(24), 2499-2507.

Geertse, T. D., Holland, R., Timmers, J. M., Paap, E., Pijnappel, R. M., Broeders, M. J., et al. (2015). Value of audits in breast cancer screening quality assurance programmes. *Eur Radiol, 25*(11), 3338-3347.

Ghate, S. V., Soo, M. S., Baker, J. A., Walsh, R., Gimenez, E. I., & Rosen, E. L. (2005). Comparison of recall and cancer detection rates for immediate versus batch interpretation of screening mammograms. *Radiology, 235*(1), 31-35.

Giess, C. S., Pourjabbar, S., Ip, I. K., Lacson, R., Alper, E., & Khorasani, R. (2017). Comparing Diagnostic Performance of Digital Breast Tomosynthesis and Full-Field Digital Mammography in a Hybrid Screening Environment. *AJR Am J Roentgenol, 209*(4), 929-934.

Given-Wilson, R., Blanks, R. (2011). Does quantity of film reading affect quality? *Clinical Radiology, 66 (2),* 67-98.

Glynn, C. G., Farria, D. M., Monsees, B. S., Salcman, J. T., Wiele, K. N., & Hildebolt, C. F. (2011). Effect of transition to digital mammography on clinical outcomes. *Radiology, 260*(3), 664-670.

Greenberg, J. S., Javitt, M. C., Katzen, J., Michael, S., & Holland, A. E. (2014). Clinical performance metrics of 3D digital breast tomosynthesis compared with 2D digital mammography for breast cancer screening in community practice. *AJR Am J Roentgenol, 203*(3), 687-693.

Gromet, M. (2008). Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. *AJR Am J Roentgenol, 190*(4), 854-859.

Hambly, N. M., McNicholas, M. M., Phelan, N., Hargaden, G. C., O'Doherty, A., & Flanagan, F. L. (2009). Comparison of digital mammography and screen-film mammography in breast cancer screening: a review in the Irish breast screening program. *AJR Am J Roentgenol, 193*(4), 1010-1018.

Hayward, J. H., Ray, K. M., Wisner, D. J., Kornak, J., Lin, W., Joe, B. N., et al. (2016). Improving Screening Mammography Outcomes Through Comparison With Multiple Prior Mammograms. *AJR Am J Roentgenol*, 1-7.

Hofvind, S., Bennett, R. L., Brisson, J., Lee, W., Pelletier, E., Flugelman, A., et al. (2016). Audit feedback on reading performance of screening mammograms: An international comparison. [Comparative Study]. *Journal of Medical Screening, 23*(3), 150-159.

Hofvind, S., Hovda, T., Holen, A. S., Lee, C. I., Albertsen, J., Bjorndal, H., et al. (2018). Digital Breast Tomosynthesis and Synthetic 2D Mammography versus Digital Mammography: Evaluation in a Population-based Screening Program. *Radiology*, 171361.

Hofvind, S., Skaane, P., Elmore, J. G., Sebuodegard, S., Hoff, S. R., & Lee, C. I. (2014). Mammographic performance in a population-based screening program: before, during, and after the transition from screen-film to full-field digital mammography. *Radiology, 272*(1), 52-62.

Hogue, J. C., Julien, M., Loisel, Y., Provencher, L., & Diorio, C. (2016). Improved detection rate of invasive breast cancers with tomosynthesis compared to 2D mammography in a screening program context. *European Journal of Cancer, 2)*, S148.

Holland, K., Sechopoulos, I., den Heeten, G., Mann, R. M., & Karssemeijer, N. (2016) Performance of breast cancer screening depends on mammographic compression. *Vol. 9699. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 183-189).

Houssami, N., Bernardi, D., Pellegrini, M., Valentini, M., Fanto, C., Ostillio, L., et al. (2017). Breast cancer detection using single-reading of breast tomosynthesis (3D-mammography) compared to double-reading of 2D-mammography: Evidence from a population-based trial. [Research Support, Non-U.S. Gov't]. *Cancer Epidemiology, 47*, 94-99.

James, J. J., & Cornford, E. J. (2009). Does computer-aided detection have a role in the arbitration of discordant double-reading opinions in a breast-screening programme? [Evaluation Studies]. *Clinical Radiology, 64*(1), 46-51.

Juel, I. M., Skaane, P., Hoff, S. R., Johannessen, G., & Hofvind, S. (2010). Screen-film mammography versus full-field digital mammography in a population-based screening program: The Sogn and Fjordane study. *Acta Radiol, 51*(9), 962-968.

Karssemeijer, N., Bluekens, A. M., Beijerinck, D., Deurenberg, J. J., Beekman, M., Visser, R., et al. (2009). Breast cancer screening results 5 years after introduction of digital mammography in a population-based screening program. *Radiology, 253*(2), 353-358.

Klompenhouwer, E. G., Duijm, L. E., Voogd, A. C., den Heeten, G. J., Strobbe, L. J., Louwman, M. W., et al. (2014). Re-attendance at biennial screening mammography following a repeated false positive recall. *Breast Cancer Research & Treatment, 145*(2), 429-437.

Klompenhouwer, E. G., Voogd, A. C., den Heeten, G. J., Strobbe, L. J., de Haan, A. F., Wauters, C. A., et al. (2015a). 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, 51*(3), 391-399.

Klompenhouwer, E. G., Voogd, A. C., den Heeten, G. J., Strobbe, L. J., Tjan-Heijnen, V. C., Broeders, M. J., et al. (2015b). Discrepant screening mammography assessments at blinded and non-blinded double reading: impact of arbitration by a third reader on screening outcome. *Eur Radiol, 25*(10), 2821-2829.

Klompenhouwer, E. G., Weber, R. J., Voogd, A. C., den Heeten, G. J., Strobbe, L. J., Broeders, M. J., et al. (2015c). Arbitration of discrepant BI-RADS 0 recalls by a third reader at screening mammography lowers recall rate but not the cancer detection rate and sensitivity at blinded and non-blinded double reading. *Breast, 24*(5), 601-607.

Lehman, C. D., Wellman, R. D., Buist, D. S., Kerlikowske, K., Tosteson, A. N., Miglioretti, D. L., et al. (2015). Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Intern Med, 175*(11), 1828-1837.

Lipasti, S., Anttila, A., & Pamilo, M. (2010). Mammographic findings of women recalled for diagnostic work-up in digital versus screen-film mammography in a population-based screening program. *Acta Radiol, 51*(5), 491-497.

Liston, J. C., & Dall, B. J. (2003). Can the NHS Breast Screening Programme afford not to double read screening mammograms? *Clinical Radiology, 58*(6), 474-477.

Lourenco, A. P., Barry-Brooks, M., Baird, G. L., Tuttle, A., & Mainiero, M. B. (2015). Changes in recall type and patient treatment following implementation of screening digital breast tomosynthesis. [Evaluation Studies]. *Radiology, 274*(2), 337-342.

McCarthy, A. M., Kontos, D., Synnestvedt, M., Tan, K. S., Heitjan, D. F., Schnall, M., et al. (2014). Screening outcomes following implementation of digital breast tomosynthesis in a general-population screening program. *J Natl Cancer Inst, 106*(11).

McDonald, E. S., McCarthy, A. M., Akhtar, A. L., Synnestvedt, M. B., Schnall, M., & Conant, E. F. (2015). Baseline Screening Mammography: Performance of Full-Field Digital Mammography Versus Digital Breast Tomosynthesis. *AJR Am J Roentgenol, 205*(5), 1143-1148.

McDonald, E. S., Oustimov, A., Weinstein, S. P., Synnestvedt, M. B., Schnall, M., & Conant, E. F. (2016). Effectiveness of Digital Breast Tomosynthesis Compared With Digital Mammography: Outcomes Analysis From 3 Years of Breast Cancer Screening.[Erratum appears in JAMA Oncol. 2016 Apr;2(4):549; PMID: 26986044]. *JAMA Oncology, 2*(6), 737-743.

Miglioretti, D. L., Gard, C. C., Carney, P. A., Onega, T. L., Buist, D. S., Sickles, E. A., et al. (2009). When radiologists perform best: the learning curve in screening mammogram interpretation. *Radiology, 253*(3), 632-640.

Mullen, L. A., Panigrahi, B., Hollada, J., Panigrahi, B., Falomo, E. T., & Harvey, S. C. (2017). Strategies for Decreasing Screening Mammography Recall Rates While Maintaining Performance Metrics. *Acad Radiol, 24*(12), 1556-1560.

Perry, N. M., Patani, N., Milner, S. E., Pinker, K., Mokbel, K., Allgood, P. C., et al. (2011). The impact of digital mammography on screening a young cohort of women for breast cancer in an urban specialist breast unit. *Eur Radiol, 21*(4), 676-682.

Posso, M. C., Puig, T., Quintana, M. J., Sola-Roca, J., & Bonfill, X. (2016). Double versus single reading of mammograms in a breast cancer screening programme: a cost-consequence analysis. *Eur Radiol, 26*(9), 3262-3271.

Powell, J. L., Hawley, J. R., Lipari, A. M., Yildiz, V. O., Erdal, B. S., & Carkaci, S. (2017). Impact of the Addition of Digital Breast Tomosynthesis (DBT) to Standard 2D Digital Screening Mammography on the Rates of Patient Recall, Cancer Detection, and Recommendations for Short-term Follow-up. *Acad Radiol, 24*(3), 302-307.

Procasco, M. (2016). Comparison of Digital Breast Tomosynthesis vs Full-Field Digital Mammography in Recall Rates and Cancer Detection Rates. *Radiol Technol, 87*(3), 349-351.

Rafferty, E. A., Rose, S. L., Miller, D. P., Durand, M. A., Conant, E. F., Copit, D. S., et al. (2017). Effect of age on breast cancer screening using tomosynthesis in combination with digital mammography. *Breast Cancer Research & Treatment, 164*(3), 659-666.

Rochman CM, Nicholson B, Peppard H, Harvey J. (no date). Reducing Recall Rates for Screening Mammography: How We Achieved Our Goal. https://www.rsna.org/uploadedFiles/RSNA/Content/Science/Quality/Storyboards/2014/Rochman-QSE012-b.pdf

Roman, R., Sala, M., Salas, D., Ascunce, N., Zubizarreta, R., Castells, X., et al. (2012). Effect of protocol-related variables and women's characteristics on the cumulative false-positive risk in breast cancer screening. *Ann Oncol, 23*(1), 104-111.

Romero Martin, S., Raya Povedano, J. L., Cara Garcia, M., Santos Romero, A. L., Pedrosa Garriguet, M., & Alvarez Benito, M. (2018). Prospective study aiming to compare 2D mammography and tomosynthesis + synthesized mammography in terms of cancer detection and recall. From double reading of 2D mammography to single reading of tomosynthesis. *Eur Radiol*.

Sala, M., Comas, M., Macia, F., Martinez, J., Casamitjana, M., & Castells, X. (2009). Implementation of digital mammography in a population-based breast cancer screening program: effect of screening round on recall rate and cancer detection. *Radiology, 252*(1), 31-39.

Sala, M., Domingo, L., Macia, F., Comas, M., Buron, A., & Castells, X. (2015). Does digital mammography suppose an advance in early diagnosis? Trends in performance indicators 6 years after digitalization. *European Radiology, 25*(3), 850-859.

Sala, M., Salas, D., Belvis, F., Sanchez, M., Ferrer, J., Ibanez, J., et al. (2011). Reduction in false-positive results after introduction of digital mammography: analysis from four population-based breast cancer screening programs in Spain. *Radiology, 258*(2), 388-395.

Salas, D., Ibanez, J., Roman, R., Cuevas, D., Sala, M., Ascunce, N., et al. (2011). Effect of start age of breast cancer screening mammography on the risk of false-positive results. *Prev Med, 53*(1-2), 76-81.

Sanchez Gomez, S., Torres Tabanera, M., Vega Bolivar, A., Sainz Miranda, M., Baroja Mazo, A., Ruiz Diaz, M., et al. (2011). Impact of a CAD system in a screen-film mammography screening program: a prospective study. *Eur J Radiol, 80*(3), e317-321.

Sankatsing, V. D. V., Fracheboud, J., de Munck, L., Broeders, M. J. M., van Ravesteyn, N. T., Heijnsdijk, E. A. M., et al. (2018). Detection and interval cancer rates during the transition from screen-film to digital mammography in population-based screening. *BMC Cancer, 18*(1), 256.

Sharpe, R. E., Jr., Venkataraman, S., Phillips, J., Dialani, V., Fein-Zachary, V. J., Prakash, S., et al. (2016). Increased Cancer Detection Rate and Variations in the Recall Rate Resulting from Implementation of 3D Digital Breast Tomosynthesis into a Population-based Screening Program. *Radiology, 278*(3), 698-706.

Shaw, C. M., Flanagan, F. L., Fenlon, H. M., & McNicholas, M. M. (2009). Consensus review of discordant findings maximizes cancer detection rate in double-reader screening mammography: Irish National Breast Screening Program experience. *Radiology, 250*(2), 354-362.

Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, Moher D, Tugwell P, Welch V, Kristjansson E, Henry DA. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017 Sep 21;358:j4008. https://amstar.ca/Amstar_Checklist.php

Smith-Bindman, R., Chu, P., Miglioretti, D. L., Quale, C., Rosenberg, R. D., Cutter, G., et al. (2005). Physician predictors of mammographic accuracy. *J Natl Cancer Inst, 97*(5), 358-367.

Tan, A., Freeman, D. H., Jr., Goodwin, J. S., & Freeman, J. L. (2006). Variation in false-positive rates of mammography reading among 1067 radiologists: a population-based assessment. *Breast Cancer Res Treat, 100*(3), 309-318.

Taylor-Phillips, S., Wallis, M. G., Duncan, A., & Gale, A. G. (2012). Use of prior mammograms in the transition to digital mammography: a performance and cost analysis. *European Journal of Radiology, 81*(1), 60-65.

Taylor-Phillips, S., Wallis, M. G., Jenkinson, D., Adekanmbi, V., Parsons, H., Dunn, J., et al. (2016). Effect of Using the Same vs Different Order for Second Readings of Screening Mammograms on Rates of Breast Cancer Detection: A Randomized Clinical Trial. *JAMA, 315*(18), 1956-1965.

Theberge, I., Chang, S. L., Vandal, N., Daigle, J. M., Guertin, M. H., Pelletier, E., et al. (2014). Radiologist interpretive volume and breast cancer screening accuracy in a Canadian organized screening program. *J Natl Cancer Inst, 106*(3), djt461.

Theberge, I., Hebert-Croteau, N., Langlois, A., Major, D., & Brisson, J. (2005). Volume of screening mammography and performance in the Quebec population-based Breast Cancer Screening Program. *CMAJ, 172*(2), 195-199.

Theberge, I., Vandal, N., Langlois, A., Pelletier, E., & Brisson, J. (2016). Detection Rate, Recall Rate, and Positive Predictive Value of Digital Compared to Screen-Film Mammography in the Quebec Population-Based Breast Cancer Screening Program. [Comparative Study]. *Canadian Association of Radiologists Journal, 67*(4), 330-338.

van Luijt, P. A., Fracheboud, J., Heijnsdijk, E. A., den Heeten, G. J., de Koning, H. J., & National Evaluation Team for Breast Cancer Screening in Netherlands Study, G. (2013). Nation-wide data on screening performance during the transition to digital mammography: observations in 6 million screens. *European Journal of Cancer, 49*(16), 3517-3525.

Van Ongeval, C., Van Steen, A., Vande Putte, G., Zanca, F., Bosmans, H., Marchal, G., et al. (2010). Does digital mammography in a decentralized breast cancer screening program lead to screening performance parameters comparable with film-screen mammography? *European Radiology, 20*(10), 2307-2314.

van Ravesteyn, N. T., Miglioretti, D. L., Stout, N. K., Lee, S. J., Schechter, C. B., Buist, D. S., et al. (2012). Tipping the balance of benefits and harms to favor screening mammography starting at age 40 years: a comparative modeling study of risk. [Research Support, N.I.H., Extramural]. *Annals of Internal Medicine, 156*(9), 609-617.

Vernacchia, F. S., & Pena, Z. G. (2009). Digital mammography: its impact on recall rates and cancer detection rates in a small community-based radiology practice. *AJR. American Journal of Roentgenology, 193*(2), 582-585.

Vinnicombe, S., Pinto Pereira, S. M., McCormack, V. A., Shiel, S., Perry, N., & Dos Santos Silva, I. M. (2009). Full-field digital versus screen-film mammography: comparison within the UK breast screening program and systematic review of published data. *Radiology, 251*(2), 347-358.

Whang, J. S., Baker, S.R., Patel, R., Luk, L., Castro, A., 3rd. (2013). The causes of medical malpractice suits against radiologists in the United States. *Radiology, 266(2),*548-554.

Yankaskas, B. C., May, R. C., Matuszewski, J., Bowling, J. M., Jarman, M. P., & Schroeder, B. F. (2011). Effect of observing change from comparison mammograms on performance of screening mammography in a large community-based population. *Radiology, 261*(3), 762-770.

Zuckerman, S. P., Conant, E. F., Keller, B. M., Maidment, A. D., Barufaldi, B., Weinstein, S. P., et al. (2016). Implementation of Synthesized Two-dimensional Mammography in a Population-based Digital Breast Tomosynthesis Screening Program. *Radiology, 281*(3), 730-736.

Appendix 1. Literature search strategy

Medline

Ovid MEDLINE(R) Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) 1946 to Present

[Search date: 19 March 2018]

#	Searches	Results
1	Breast Neoplasms/di, dg [Diagnosis, Diagnostic Imaging]	48489
2	(Breast adj3 tumo?r).mp.	16190
3	(Breast adj3 neoplasm).mp.	737
4	(Breast adj3 cancer).mp.	241751
5	(Breast adj3 carcinoma).mp.	42884
6	(Mammary adj3 tumo?r).mp.	10802
7	(Mammary adj3 neoplasm).mp.	62
8	(Mammary adj3 cancer).mp.	3707
9	(Mammary adj3 carcinoma).mp.	7067
10	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9	299036
11	Mass Screening/mt [Methods]	27829
12	"Early Detection of Cancer"/	17615
13	mass screening.mp.	95215
14	screening program*.mp.	25794
15	(early detection adj2 cancer).mp.	19492
16	11 or 12 or 13 or 14 or 15	124825
17	false positive reactions/	26404
18	diagnostic errors/	35076
19	false positive*.mp.	67099
20	diagnostic error*.mp.	36828
21	abnormal call*.mp.	39
22	recall rate*.mp.	936
23	17 or 18 or 19 or 20 or 21 or 22	102701
24	radiographic image interpretation, computer-assisted/	12727
25	radiographic image enhancement/	18025
26	breast imaging.mp.	3285
27	tomosynthesis.mp.	1237
28	(film adj1 screen mammography).mp.	126
29	film-screen mammography.mp.	126
30	screen-film.mp.	1064
31	digital mammography.mp.	1581
32	computed radiography.mp.	1025
33	digital radiography.mp.	1849
34	computer assisted system*.mp.	304
35	computer-aided detection system*.mp.	137
36	24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35	33108

37	MAMMOGRAPHY/st [Standards]	1128
38	*RADIOLOGY/st [Standards]	1315
39	RADIOLOGY/ma [Manpower]	802
40	Double read*.mp.	346
41	reading volume*.mp.	19
42	(quality adj2 assurance).mp.	67839
43	(quality adj2 improvement).mp.	43267
44	(quality adj2 report*).mp.	8333
45	(quality adj2 control).mp.	75003
46	(quality adj2 management).mp.	19107
47	quality assurance practice*.mp.	92
48	program performance.mp.	448
49	37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46 or 47 or 48	195728
50	professional competence/	22817
51	defensive medicine/	1188
52	Liability, legal/	15176
53	malpractice/	26992
54	50 or 51 or 52 or 53	60569
55	radiologist/	339
56	54 and 55	6
57	Radiologists/ed, lj, sn [Education, Legislation & Jurisprudence, Statistics & Numerical Data]	86
58	(Radiologist* adj3 detection measure*).mp.	1
59	(radiologist adj3 interpretive efficiency).mp.	1
60	(radiologist adj3 demographic*).mp.	13
61	(radiologist adj3 training).mp.	74
62	(radiologist adj3 education).mp.	19
63	(radiologist adj3 competence).mp.	3
64	(radiologist adj3 experience).mp.	132
65	(radiologist adj3 gender).mp.	5
66	(radiologist adj3 values).mp.	18
67	(radiologist* and litigation).mp.	69
68	57 or 58 or 59 or 60 or 61 or 62 or 63 or 64 or 65 or 66 or 67	415
69	56 or 68	416
70	36 or 49 or 69	227694
71	10 and 16 and 23 and 70	431
72	limit 71 to (english language and yr="2003 -Current")	353

Embase

Embase Classic+Embase 1947 to 2018 March 16
[Search date: 19 March 2018]

#	Searches	Results
1	exp breast tumor/di [Diagnosis]	59049
2	(Breast adj3 tumo?r).mp.	100177
3	(Breast adj3 neoplasm).mp.	1873
4	(Breast adj3 cancer).mp.	442517
5	(Breast adj3 carcinoma).mp.	87234
6	(Mammary adj3 tumo?r).mp.	14101
7	(Mammary adj3 neoplasm).mp.	733
8	(Mammary adj3 cancer).mp.	5949
9	(Mammary adj3 carcinoma).mp.	10615
10	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9	530222
11	mass screening/	54937
12	early cancer diagnosis/	2793
13	mass screening.mp.	58697
14	screening program*.mp.	36467
15	(early detection adj2 cancer).mp.	3384
16	11 or 12 or 13 or 14 or 15	93963
17	false positive result/	22827
18	diagnostic error/	53654
19	false positive*.mp.	81436
20	diagnostic error*.mp.	55537
21	abnormal call*.mp.	48
22	recall rate*.mp.	1182
23	17 or 18 or 19 or 20 or 21 or 22	135080
24	computer assisted radiography/ or digital mammography/	3204
25	image enhancement/	26407
26	breast imaging.mp.	4375
27	tomosynthesis.mp.	1585
28	(film adj1 screen mammography).mp.	156
29	film-screen mammography.mp.	155
30	screen-film.mp.	1244
31	digital mammography.mp.	2960
32	computed radiography.mp.	1252
33	digital radiography.mp.	5051
34	computer assisted system*.mp.	370
35	computer-aided detection system*.mp.	180
36	24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35	41280
37	Double read*.mp.	487
38	reading volume*.mp.	25
39	(quality adj2 assurance).mp.	35620
40	(quality adj2 improvement).mp.	49936

41	(quality adj2 report*).mp.	11758
42	(quality adj2 control).mp.	193297
43	(quality adj2 management).mp.	55437
44	quality assurance practice*.mp.	130
45	program performance.mp.	521
46	37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45	289062
47	professional competence/	29589
48	defensive medicine/	309
49	legal liability/	16119
50	malpractice/	32640
51	47 or 48 or 49 or 50	73787
52	radiologist/	37640
53	51 and 52	562
54	(radiologist* adj3 detection measure*).mp.	1
55	(radiologist adj3 interpretive efficiency).mp.	1
56	(radiologist adj3 demographic*).mp.	28
57	(radiologist adj3 training).mp.	112
58	(radiologist adj3 education).mp.	33
59	(radiologist adj3 competence).mp.	8
60	(radiologist adj3 experience).mp.	213
61	(radiologist adj3 gender).mp.	8
62	(radiologist adj3 values).mp.	26
63	(radiologist* and litigation).mp.	118
64	54 or 55 or 56 or 57 or 58 or 59 or 60 or 61 or 62 or 63	541
65	53 or 64	1060
66	36 or 46 or 65	330015
67	10 and 16 and 23 and 66	241
68	limit 67 to (english language and yr="2003 -Current")	179

Scopus

[Search date: 19 March 2018]

#	Search	Results
1	(TITLE-ABS-KEY (breast W/3 tumor) OR TITLE-ABS-KEY (breast W/3 neoplasm) OR TITLE-ABS-KEY (breast W/3 cancer) OR TITLE-ABS-KEY (breast W/3 carcinoma) OR TITLE-ABS-KEY (mammary W/3 tumor) OR TITLE-ABS-KEY (mammary W/3 neoplasm) OR TITLE-ABS-KEY (mammary W/3 cancer) OR TITLE-ABS-KEY (mammary W/3 carcinoma))	494,202
2	(TITLE-ABS-KEY (mass W/1 screening) OR TITLE-ABS-KEY (screening W/1 program) OR TITLE-ABS-KEY (early W/2 detection W/2 cancer) OR TITLE-ABS-KEY ({early detection of cancer}))	145,367
3	(TITLE-ABS-KEY (diagnostic PRE/1 error) OR TITLE-ABS-KEY (false PRE/1 positive) OR TITLE-ABS-KEY (abnormal PRE/1 call) OR TITLE-ABS-KEY (recall PRE/1 rate))	159,300
4	(TITLE-ABS-KEY ({computer-assisted radiographic image interpretation}) OR TITLE-ABS-KEY ({radiographic image enhancement}) OR TITLE-ABS-KEY (breast W/2 imaging) OR TITLE-ABS-KEY (tomosynthesis) OR TITLE-ABS-KEY (film PRE/1 screen PRE/1 mammography) OR TITLE-ABS-KEY (screen PRE/1 film) OR TITLE-ABS-KEY (digital PRE/1 mammography) OR TITLE-ABS-KEY (computed PRE/1 radiography) OR TITLE-ABS-KEY (digital PRE/1 radiography) OR TITLE-ABS-KEY (computer PRE/1 assisted PRE/1 system) OR TITLE-ABS-KEY ({computer-aided detection system}) OR TITLE-ABS-KEY ({computer aided detection system}))	40,096
5	(TITLE-ABS-KEY (double PRE/1 read*) OR TITLE-ABS-KEY (reading PRE/1 volume) OR TITLE-ABS-KEY (quality W/2 assurance) OR TITLE-ABS-KEY (quality W/2 improvement) OR TITLE-ABS-KEY (quality W/2 report*) OR TITLE-ABS-KEY (quality W/2 control) OR TITLE-ABS-KEY (quality W/2 management) OR TITLE-ABS-KEY (program PRE/1 performance))	611,593
6	(TITLE-ABS-KEY (radiologist* W/3 detection W/3 measure*) OR TITLE-ABS-KEY (radiologist W/3 interpretive W/3 efficiency) OR TITLE-ABS-KEY (radiologist W/3 demographic*) OR TITLE-ABS-KEY (radiologist W/3 training) OR TITLE-ABS-KEY (radiologist W/3 education) OR TITLE-ABS-KEY (radiologist W/3 competence) OR TITLE-ABS-KEY (radiologist W/3 experience) OR TITLE-ABS-KEY (radiologist W/3 gender) OR TITLE-ABS-KEY (radiologist W/3 values) OR TITLE-ABS-KEY ((radiologist AND litigation)) OR TITLE-ABS-KEY ((radiologist AND malpractice)))	2,153
7	4 OR 5 OR 6	651,476
8	1 AND 2 AND 3 AND 7	559
9	Limit to year 2003-current	480
10	Limit to English language	448

Cochrane Database of Systematic Reviews

[Search date: 21 March 2018]

#	Searches	Results
1	(Breast adj3 tumo?r).mp.	74
2	(Breast adj3 neoplasm).mp.	57
3	(Breast adj3 cancer).mp.	434
4	(Breast adj3 carcinoma).mp.	62
5	(Mammary adj3 tumo?r).mp.	5
6	(Mammary adj3 neoplasm).mp.	11
7	(Mammary adj3 carcinoma).mp.	9
8	1 or 2 or 3 or 4 or 5 or 6 or 7	443
9	screening.mp.	3861
10	early detection.mp.	213
11	9 or 10	3912
12	false positive*.mp.	489
13	diagnostic error*.mp.	25
14	abnormal call*.mp.	0
15	recall rate*.mp.	2
16	12 or 13 or 14 or 15	501
17	breast imaging.mp.	3
18	tomosynthesis.mp.	0
19	(film adj1 screen mammography).mp.	1
20	film-screen mammography.mp.	1
21	screen-film.mp.	1
22	digital mammography.mp.	0
23	computed radiography.mp.	2
24	digital radiography.mp.	2
25	computer assisted system*.mp.	0
26	computer-aided detection system*.mp.	0
27	17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26	7
28	Double read*.mp.	0
29	reading volume*.mp.	0
30	(quality adj2 assurance).mp.	107
31	(quality adj2 improvement).mp.	705
32	(quality adj2 report*).mp.	2413
33	(quality adj2 control).mp.	394
34	(quality adj2 management).mp.	77
35	quality assurance practice*.mp.	0
36	program performance.mp.	2
37	28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36	3055
38	(radiologist* adj3 detection measure*).mp.	0
39	(radiologist adj3 interpretive efficiency).mp.	0
40	(radiologist adj3 demographic*).mp.	0
41	(radiologist adj3 training).mp.	3
42	(radiologist adj3 education).mp.	0

43	(radiologist adj3 competence).mp.	0
44	(radiologist adj3 experience).mp.	7
45	(radiologist adj3 gender).mp.	0
46	(radiologist adj3 values).mp.	0
47	(radiologist* and litigation).mp.	0
48	38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46 or 47	10
49	27 or 37 or 48	3067
50	8 and 11 and 16 and 49	13
51	limit 50 to last 15 years	11

Appendix 2. Identification of start date for original studies from relevant systematic reviews.

Table A1. Technology

Reference	Comprehensiveness	Use for Selection of Publication Date?
	Screen-film vs. Digital Mammography	
Irwig L, Houssami N, van Vliet C. New technologies in screening for breast cancer: a systematic review of their accuracy. Br J Cancer. 2004;90(11):2118-2122 **[Systematic Review - Critically Low Quality]**	**Methods** • "MEDLINE was searched from 1966 to December 2002" • "The search was extended by examining references given in relevant primary studies and review articles, contact with content experts, and targeted further MEDLINE searches, for example on authors of earlier studies." **References Included** • Full-field digital mammography (FFDM) vs. Conventional mammography (N = 2 articles describing 1 study)	**Decision** • No **Justification** • Systematic Review search date does not cover publication dates of interest
Elmore, J. G., Armstrong, K., Lehman, C. D., & Fletcher, S. W. (2005). Screening for breast cancer. JAMA, 293(10), 1245-1256. **[Systematic Review - Critically Low Quality]**	**Methods** • "searches of MEDLINE, The Cochrane Library, the National Guideline Clearinghouse Web site, the US Preventive Services Task Force recommendations and reviews,[5,13] and the International Agency for Research on Cancer Handbook of Cancer Prevention (IARC)[4] were performed to identify English-language articles about breast cancer screening" • "The bibliographies of retrieved articles were also scanned to retrieve additional relevant articles." **References Included**	**Decision** • No **Justification** • More recent systematic review available

Reference	Comprehensiveness	Use for Selection of Publication Date?
	- Full-field digital mammography vs. Screen-film mammography (N = 3 community-based studies). Publication dates range from 2002 – 2004	
Vinnicombe, S., Pinto Pereira, S. M., McCormack, V. A., Shiel, S., Perry, N., & Dos Santos Silva, I. M. (2009). Full-field digital versus screen-film mammography: comparison within the UK breast screening program and systematic review of published data. Radiology, 251(2), 347-358. **[Systematic Review - - Critically Low Quality]**	**Methods** - "search was conducted by using PubMed, MEDLINE, and EMBASE to identify studies published in English-language journals between January 1, 2000, and February 29, 2008 (inclusive) that compared FFDM to SFM in terms of their process indicators" - "Reference lists within relevant articles and reviews were searched to identify further publications." **References Included** - Screen-film mammography vs. Full-field digital mammography (N = 8 studies)	**Decision** - No **Justification** - More recent systematic review available
Iared, W., Shigueoka, D. C., Torloni, M. R., Velloni, F. G., Ajzen, S. A., Atallah, A. N., et al. (2011). Comparative evaluation of digital mammography and film mammography: Systematic review and meta-analysis. Sao Paulo Medical Journal, 129(4), 250-260. **[Systematic Review - Critically Low Quality]**	**Methods** - "search strategy involved searching four electronic databases (Medline via PubMed, Embase, Lilacs and Scopus) for articles on the topics of digital and film mammography that had been published up to September 2009. The bibliographic references of the studies included were checked in order to search for additional potentially relevant citations." **References Included** - Film mammography vs. digital mammography (N = 11 studies)	**Decision** - Yes **Justification** - Most recent systematic review

Reference	Comprehensiveness	Use for Selection of Publication Date?
Mohd Norsuddin, N., Reed, W., Mello-Thoms, C., & Lewis, S. J. (2015). Understanding recall rates in screening mammography: A conceptual framework review of the literature. Radiography, 21(4), 334-341. **[Review-Critically Low Quality]**	**N/A**	**N/A**
Le, M. T., Mothersill, C. E., Seymour, C. B., & McNeill, F. E. (2016). Is the false-positive rate in mammography in North America too high? [Review]. British Journal of Radiology, 89(1065), 20160045. **[Review]**	**N/A**	**N/A**
		START DATE FOR SEARCH OF ORIGINAL RESEARCH: 2009
	Computer-Aided Detection Systems	
Irwig L, Houssami N, van Vliet C. New technologies in screening for breast cancer: a systematic review of their accuracy. Br J Cancer. 2004;90(11):2118-2122 **[Systematic Review - Critically Low Quality]**	**Methods** • "MEDLINE was searched from 1966 to December 2002" • "The search was extended by examining references given in relevant primary studies and review articles, contact with content experts, and targeted further MEDLINE searches, for example on authors of earlier studies." **References Included** • N = 4 articles describing 3 studies on CAD	**Decision** • No **Justification** • Systematic review search date does not cover publication dates of interest

Reference	Comprehensiveness	Use for Selection of Publication Date?
Elmore, J. G., Armstrong, K., Lehman, C. D., & Fletcher, S. W. (2005). Screening for breast cancer. JAMA, 293(10), 1245-1256. **[Systematic Review - Critically Low Quality]**	**Methods** - "searches of MEDLINE, The Cochrane Library, the National Guideline Clearinghouse Web site, the US Preventive Services Task Force recommendations and reviews,[5,13] and the International Agency for Research on Cancer Handbook of Cancer Prevention (IARC)[4] were performed to identify English-language articles about breast cancer screening" - "The bibliographies of retrieved articles were also scanned to retrieve additional relevant articles." **References Included** - No CAD vs. CAD (N = 2 studies). Publication dates range from 2001 - 2004	**Decision** - No **Justification** - More recent systematic review available
Taylor, P., & Potts, H. W. (2008). Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer, 44(6), 798-807. **[Systematic Review – Critically Low Quality]**	**Methods** - "The NLH PubMed database was searched…. Google Scholar, Biotech, CINAHL, Embase, HMIC, Pyschinfo, Web of Science and Science Direct were searched…. The online catalogue of the British Library and recent proceedings of relevant conferences were searched. A previous systematic review of double reading was identified and its references were checked,[3] as were references in retrieved papers." **References Included** - Single reading with CAD vs. single reading (N = 10 studies). Publication dates range from 2001 – 2008	**Decision** 1. Yes **Justification** - Most recent systematic review available on single reading with CAD vs. single reading

Reference	Comprehensiveness	Use for Selection of Publication Date?
Noble, M., Bruening, W., Uhl, S., & Schoelles, K. (2009). Computer-aided detection mammography for breast cancer screening: systematic review and meta-analysis. Archives of Gynecology & Obstetrics, 279(6), 881-890. **[Systematic Review - Moderate Quality]**	**Methods** • "searched seventeen databases including MEDLINE, EMBASE, and the Cochrane Library though September 25, 2008, and we hand-searched the bibliographies/reference lists from peer-reviewed and gray literature (i.e. reports and studies produced by local government agencies, private organizations, educational facilities, and corporations) to identify clinical studies not identified by the electronic searches." **References Included** • N = 5 studies on CAD	**Decision** • No **Justification** • Unclear comparators (single/double reading)
Azavedo E, Zackrisson S, Mejàre I, Heibert Arnlind M. Is single reading with computer-aided detection (CAD) as good as double reading in mammography screening? A systematic review. BMC Med Imaging. 2012 Jul 24;12:22. **[Systematic Review – Moderate Quality]**	**Methods** • "The electronic literature search included the databases PubMed, EMBASE, and The Cochrane Library from 1950 to November 2011. All Western European languages were accepted." • "Hand search and grey literature did not result in any additional articles." **References Included** • Single reading + CAD vs. double reading (N = 4)	**Decision** • Yes **Justification** • Most recent systematic review on single reading with CAD vs. double reading
Astley SM, Gilbert FJ. Computer-aided detection in mammography. Clin Radiol. 2004;59(5):390-399 **[Review]**	**N/A**	**N/A**

Reference	Comprehensiveness	Use for Selection of Publication Date?
Houssami, N., Given-Wilson, R., & Ciatto, S. (2009). Early detection of breast cancer: overview of the evidence on computer-aided detection in mammography screening. Journal of Medical Imaging & Radiation Oncology, 53(2), 171-176. **[Review]**	**N/A**	**N/A**

START DATE FOR SEARCH OF ORIGINAL RESEARCH: 2008 for single reading + CAD vs. single reading; 2011 for single reading + CAD vs. double reading

Tomosynthesis

Svahn, T. M., Macaskill, P., & Houssami, N. (2015). Radiologists' interpretive efficiency and variability in true- and false-positive detection when screen-reading with tomosynthesis (3D-mammography) relative to standard mammography in population screening. Breast, 24(6), 687-693. **[Systematic Review - Critically Low Quality]**	**Methods** - "To identify all potentially eligible primary studies we systematically searched the literature; replicating a previously published systematic search [12] in January 2015, and further updated the search at week 1 July 2015. The search strategy consisted of a Medline search …and also contact with content experts." **References Included** - 2D/3D vs. 2D (N = 3 studies / 4 publications)	**Decision** - No **Justification** - Systematic reviews of higher quality are available
Hodgson, R., Heywang-Kobrunner, S. H., Harvey, S. C., Edwards, M., Shaikh, J., Arber, M., et al. (2016). Systematic review of 3D mammography for breast cancer screening.	**Methods** - "This systematic review was carried out according to the systematic review guidance provided in the Cochrane Handbooks …The searches were performed and concluded in October 2014."	**Decision** - Yes **Justification** - Systematic review is of moderate quality

Reference	Comprehensiveness	Use for Selection of Publication Date?
[Review]. Breast, 27, 52-61. **[Systematic Review – Moderate Quality]**	• “Reference lists of relevant papers retrieved by the searches were scanned for potentially eligible studies. Systematic reviews identified by the searches were checked for additional reported research not retrieved by the database searches. Citation searches were carried out on identified records.” **References Included** • FFDM vs. DBT+FFDM (N = 5 studies / 16 reports)	
Nelson, H. D., Pappas, M., Cantor, A., Griffin, J., Daeges, M., & Humphrey, L. (2016). Harms of Breast Cancer Screening: Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation. Annals of Internal Medicine, 164(4), 256-267. **[Systematic Review – Moderate Quality]**	**Methods** • “A research librarian conducted electronic searches of the Cochrane Central Register of Controlled Trials, the Cochrane Database of Systematic Reviews, and Ovid MEDLINE through December 2014 for relevant studies and systematic reviews. Searches were supplemented by references identified from additional sources, including reference lists and experts. Studies of harms included in the previous systematic review for the USPSTF (2, 3) were also included.” **References Included** • DM vs. DM+tomosynthesis (N = 5 Studies)	**Decision** • Yes **Justification** • Systematic review is of moderate quality
Pozz, A., Corte, A. D., Lakis, M. A., & Jeong, H. (2016). Digital Breast Tomosynthesis in Addition to Conventional 2D Mammography Reduces Recall Rates and is CostEffective. Asian Pacific Journal of Cancer Prevention:	**Methods** • “A comprehensive systematic review was conducted independently by all three authors using search terms such as tomosynthesis, breast imaging, 3D-mammography. PubMed, Medline, Google Scholar, Ovid, and Cochrane data search engines were utilized from	**Decision** • No **Justification** • Systematic Reviews of higher quality are available

Reference	Comprehensiveness	Use for Selection of Publication Date?
Apjcp, 17(7), 3521-3526. **[Systematic Review - Critically Low Quality]**	inception until April 2016. The authors then manually scrutinized reference lists in the recovered articles and relevant abstracts from scientific meetings to identify any further articles." **References Included** • DBT vs. digital mammography (N=3 references) • DBT+DM vs. DM (N=10/11 references report on outcomes of interest)	
Cole, E. B., & Pisano, E. D. (2016). Tomosynthesis for breast cancer screening. [Article]. Clinical Imaging, 40(2), 283-287. **[Review]**	**N/A**	**N/A**
Gilbert, F. J., Tucker, L., & Young, K. C. (2016). Digital breast tomosynthesis (DBT): a review of the evidence for use as a screening tool. Clinical Radiology, 71(2), 141-150. **[Review]**	**N/A**	**N/A**
Vedantham, S., Karellas, A., Vijayaraghavan, G. R., & Kopans, D. B. (2015). Digital Breast Tomosynthesis: State of the Art. [Comparative Study]. Radiology, 277(3), 663-684. **[Review]**	**N/A**	**N/A**

Reference	Comprehensiveness	Use for Selection of Publication Date?
Skaane, P. (2017). Breast cancer screening with digital breast tomosynthesis. Breast Cancer, 24(1), 32-41. **[Review]**	**N/A**	**N/A**

START DATE FOR SEARCH OF ORIGINAL RESEARCH: 2014

Table A2. Quality assurance practices

Reference	Comprehensiveness	Use for Selection of Publication Date?
Reading Volume		
Mohd Norsuddin, N., Reed, W., Mello-Thoms, C., & Lewis, S. J. (2015). Understanding recall rates in screening mammography: A conceptual framework review of the literature. Radiography, 21(4), 334-341. **[Review - Critically Low Quality]**	**N/A**	**N/A**
Le, M. T., Mothersill, C. E., Seymour, C. B., & McNeill, F. E. (2016). Is the false-positive rate in mammography in North America too high? [Review]. British Journal of Radiology, 89(1065), 20160045. **[Review]**	**N/A**	**N/A**
		START DATE FOR SEARCH OF ORIGINAL RESEARCH: 2003
Double Reading		
Taylor, P., & Potts, H. W. (2008). Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer, 44(6), 798-807. **[Systematic Review Critically Low Quality]**	**Methods** • ""The NLH PubMed database was searched…. Google Scholar, Biotech, CINAHL, Embase, HMIC, Pyschinfo, Web of Science and Science Direct were searched…. The online catalogue of the British Library and recent proceedings of relevant conferences were searched. A previous systematic review of double reading was identified and its references were	**Decision** • No **Justification:** • Inclusion criteria includes second reader not being a radiologist

Reference	Comprehensiveness	Use for Selection of Publication Date?
	checked,[3] as were references in retrieved papers." **References Included** - Double reading vs. single reading (N = 17). Publication dates range from 1991 – 2008	
Pow, R. E., Mello-Thoms, C., & Brennan, P. (2016). Evaluation of the effect of double reporting on test accuracy in screening and diagnostic imaging studies: A review of the evidence. J Med Imaging Radiat Oncol, 60(3), 306-314. **[Systematic Review - Critically Low Quality]**	**Methods** - "A broad literature search was carried out between June and November 2015, using PubMed, CINAHL, EMBASE, Web of Science and Google Scholar. Other sources reviewed included national cancer-screening guidelines and reference lists of retrieved articles." **References Included** - Screening mammography (N = 22 studies)	**Decision** - No **Justification** - Unclear how many studies reported on recall rate
Hackney, L., Szczepura, A., Moody, L., & Whiteman, B. (2017). Review of the evidence on the use of arbitration or consensus within breast screening: A systematic scoping review. Radiography (Lond), 23(2), 171-176. **[Systematic Review – Moderate Quality]**	**Methods** - "Literature searches of PubMed, Medline, CINAHL, EMBASE, Scopus, Web of Science and the Cochrane Library were supplemented by a broad Google scholar web search. Hand searching of key peer-reviewed breast and radiology journals, a manual search of reference lists and key author searching was undertaken. Grey literature was sourced by hand searching of conference proceedings and doctoral theses. Personal contact with experts internationally was also undertaken in locating relevant literature." **References Included** - N = 26 studies	**Decision** - No **Justification** - Focus is on arbitration and consensus in double reading

Reference	Comprehensiveness	Use for Selection of Publication Date?
Posso, M., Puig, T., Carles, M., Rue, M., Canelo-Aybar, C., & Bonfill, X. (2017). Effectiveness and cost-effectiveness of double reading in digital mammography screening: A systematic review and meta-analysis. European Journal of Radiology, 96, 40-49. **[Systematic Review – Moderate Quality]**	**Methods** • "Databases were searched from 1st January 1990 to 20th February 2017, including Medline, EMBASE, and the Cochrane Library" • "We included studies we deemed as relevant based on our previous experience, and hand searched the bibliography of the included studies" **References Included** • Double reading vs. single reading of digital mammograms (N= 2 studies)	**Decision** • No **Justification** • Focus is on false positives instead of on recall rate
Mohd Norsuddin, N., Reed, W., Mello-Thoms, C., & Lewis, S. J. (2015). Understanding recall rates in screening mammography: A conceptual framework review of the literature. Radiography, 21(4), 334-341. **[Review- Critically Low quality]**	**N/A**	**N/A**
Le, M. T., Mothersill, C. E., Seymour, C. B., & McNeill, F. E. (2016). Is the false-positive rate in mammography in North America too high? [Review]. British Journal of Radiology, 89(1065), 20160045. **[Review]**	**N/A**	**N/A**

Reference	Comprehensiveness	Use for Selection of Publication Date?
		START DATE FOR SEARCH OF ORIGINAL RESEARCH: 2003
Audit/Performance Feedback		
Soh, B. P., Lee, W., Kench, P. L., Reed, W. M., McEntee, M. F., Poulos, A., et al. (2012). Assessing reader performance in radiology, an imperfect science: lessons from breast screening. [Review]. Clinical Radiology, 67(7), 623-628 **[Review]**	N/A	N/A
		START DATE FOR SEARCH OF ORIGINAL RESEARCH: 2003
Comparison with Prior Mammograms		
None Identified		
		START DATE FOR SEARCH OF ORIGINAL RESEARCH: 2003
Number of Mammographic Views		
Le, M. T., Mothersill, C. E., Seymour, C. B., & McNeill, F. E. (2016). Is the false-positive rate in mammography in North America too high? [Review]. British Journal of Radiology, 89(1065), 20160045. **[Review]**	N/A	N/A
		START DATE FOR SEARCH OF ORIGINAL RESEARCH: 2003

Table A3. Radiologist characteristics

Reference	Comprehensiveness	Use for Selection of Publication Date?
Training, Education, and Experience		
van den Biggelaar, F. J., Nelemans, P. J., & Flobbe, K. (2008). Performance of radiographers in mammogram interpretation: a systematic review. Breast, 17(1), 85-90. **[Systematic Review - Critically Low Quality]**	**N/A**	**Decision** • N/A **Justification** • Comparison of performance of non-radiologists with that of radiologists in mammogram interpretation
Mohd Norsuddin, N., Reed, W., Mello-Thoms, C., & Lewis, S. J. (2015). Understanding recall rates in screening mammography: A conceptual framework review of the literature. Radiography, 21(4), 334-341. **[Review - Critically Low Quality]**	**N/A**	**N/A**
		START DATE FOR SEARCH OF ORIGINAL RESEARCH: 2003
Age and Gender		
None Identified		
		START DATE FOR SEARCH OF ORIGINAL RESEARCH: 2003

	Litigation Concerns	
Le, M. T., Mothersill, C. E., Seymour, C. B., & McNeill, F. E. (2016). Is the false-positive rate in mammography in North America too high? [Review]. British Journal of Radiology, 89(1065), 20160045. **[Review]**	**N/A**	**N/A**
		START DATE FOR SEARCH OF ORIGINAL RESEARCH: 2003

Appendix 3. Data from review articles

Table A4. Technology

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
Screen-film vs. Digital Mammography			
Irwig L, Houssami N, van Vliet C. New technologies in screening for breast cancer: a systematic review of their accuracy. Br J Cancer. 2004;90(11):2118-2122 [**Systematic Review - Critically Low Quality**]	• Full-field digital mammography (FFDM) vs. Conventional mammography	• "MEDLINE was searched from 1966 to December 2002" • "The search was extended by examining references given in relevant primary studies and review articles, contact with content experts, and targeted further MEDLINE searches, for example on authors of earlier studies." • Full-field digital mammography (FFDM) vs. Conventional mammography (N = 1 article)	**Results** • One study reports on recall rates in FFDM vs. conventional mammography. This study demonstrates a significantly lower recall rates in FFDM vs. conventional mammography (11.8% vs. 14.9%, P<0.001). FFDM has lower overall sensitivity (64.3%) than conventional mammography (78.6%) but identifies 21.4% additional cancers that are not identified on conventional mammography. **Conclusion** • One study of FFDM suggests that it may identify some cancers not identified on conventional mammography and may result in a lower recall rate. The evidence is currently insufficient to support the use of any of these new technologies in population screening, but would support further evaluation
Elmore, J. G., Armstrong, K., Lehman, C. D., & Fletcher, S. W. (2005). Screening for breast cancer. JAMA, 293(10), 1245-1256. [**Systematic Review - Critically Low Quality**]	• Full-field digital mammography (FFDM) vs. Screen-film mammography	• "searches of MEDLINE, The Cochrane Library, the National Guideline Clearinghouse Web site, the US Preventive Services Task Force recommendations and reviews,[5,13] and the International Agency for Research on Cancer Handbook of Cancer Prevention (IARC)[4] were performed to identify English-language articles about breast cancer screening"	**Results** • "Two studies found the sensitivity of full-field digital mammography (64% and 74%) to be less than that of screen-film mammography (79%, 90%), but these studies had a small number of women with breast cancer (42 and 31, respectively) and the display systems and experience of radiologists may have improved since these studies…A larger randomized study reported similar cancer detection rates (per all screened), with higher recall rates for full-field digital mammography…" **Conclusion** • "Studies comparing full-field digital mammography to screen film have not shown statistically significant differences in cancer detection while the impact on recall rates…was unclear".

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
		• "The bibliographies of retrieved articles were also scanned to retrieve additional relevant articles." • Full-field digital mammography vs. Screen-film mammography (N = 3 community-based studies). Publication dates range from 2002 – 2004	
Vinnicombe, S., Pinto Pereira, S. M., McCormack, V. A., Shiel, S., Perry, N., & Dos Santos Silva, I. M. (2009). Full-field digital versus screen-film mammography: comparison within the UK breast screening program and systematic review of published data. Radiology, 251(2), 347-358. [**Systematic Review - - Critically Low Quality**]	• Full-field digital mammography (FFDM) using hardcopy image reading vs. conventional screen-film mammography	• "search was conducted by using PubMed, MEDLINE, and EMBASE to identify studies published in English-language journals between January 1, 2000, and February 29, 2008 (inclusive) that compared FFDM to SFM in terms of their process indicators" • "Reference lists within relevant articles and reviews were searched to identify further publications." • Screen-film mammography vs. Full-field digital mammography (N = 8 studies)	**Results** • "Recall rates varied greatly between studies, with much higher rates in the United States than in the European and Japanese studies… There was marked between-study heterogeneity in differences in recall rates between modalities (I2=94%), with some studies showing significantly lower and others significantly higher recall rates for FFDM; thus, pooled estimates could not be calculated. Similarly, there was marked between-study type heterogeneity in modality differences in the PPV of an abnormal mammogram (I2=100%), with only cohort studies showing a higher PPV for FFDM… and, hence, no pooled estimates were calculated. Some studies…presented various recall and PPV estimates for different definitions of abnormal mammograms (eg, before or after consensus meetings) and detected cancers (eg, at initial screening only or during follow-up), but these alternative estimates did not affect the findings…" • "The overall pooled estimate was consistent with FFDM having a higher detection rate than SFM (pooled FFDM-SFM difference, 0.04 [95% CI: -0.03, 0.11] per 100 screening mammograms, equivalent to FFDM depicting an extra four cases of breast cancer per every 10 000 screening mammograms), but with evidence of some heterogeneity between study types (I^{2}= 40%)" **Conclusion** • "FFDM with hardcopy image reading performed as well as SFM in terms of process indicators; the meta-analysis was consistent with FFDM yielding detection rates at least as high as those for SFM.
Iared, W., Shigueoka, D. C., Torloni, M. R., Velloni, F. G., Ajzen, S. A., Atallah,	• Film mammography vs. digital mammography	• "search strategy involved searching four electronic databases (Medline via PubMed, Embase, Lilacs and	**Results** • "There was great heterogeneity among the studies with regard to the patient recall rate (I^2 = 96%), even when they were analyzed according to study design (for cohort studies, I^2 = 95%; for paired studies, I^2 = 93%). The meta-

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
A. N., et al. (2011). Comparative evaluation of digital mammography and film mammography: Systematic review and meta-analysis. Sao Paulo Medical Journal, 129(4), 250-260. **[Systematic Review - Critically Low Quality]**		Scopus) for articles on the topics of digital and film mammography that had been published up to September 2009. The bibliographic references of the studies included were checked in order to search for additional potentially relevant citations.” • Film mammography vs. digital mammography (N = 11 studies)	analysis did not identify any significant difference between the two methods with regard to the patient recall rate: RR = 1.07; 95% CI = 0.94-1.22; I^2 = 96%). However, the RCT revealed a significant difference, with higher recall rates among patients who underwent digital mammography (RR = 1.69; 95% CI = 1.46-1.96).” • “The results showed homogeneity in terms of the cancer detection rate. The cancer detection rate was significantly higher among patients who underwent digital mammography. Based on the combination of data from the 11 studies included in this systematic review, the average relative-risk estimate for cancer detection among patients who underwent digital mammography was 1.17 (95% confidence interval, CI = 1.06-1.29; I^2 = 19%), in relation to film mammography.” **Conclusion** • “The cancer detection rates using digital mammography are slightly higher than the rates using film mammography. There are no significant differences in recall rates between film and digital mammography.”
Mohd Norsuddin, N., Reed, W., Mello-Thoms, C., & Lewis, S. J. (2015). Understanding recall rates in screening mammography: A conceptual framework review of the literature. Radiography, 21(4), 334-341. [**Narrative review**]	• Full-field digital mammography (FFDM) vs. Screen-film mammography (SFM)	• “The search of the literature was conducted in MEDLINE, CINAHL (EbSCOhost), SPIE library,Web of Science, PubMed, Scopus databases and Google Scholar. No specific year of publication was imposed in this search however, we prioritised studies from 2000 onwards which were likely to capture current imaging modalities in screening mammography.” • Note: although the authors report on databases searched and keywords used, they do not classify their review as systematic.	**Results** • “A clinical trial by Lewin et al.... in the Colorado-Massachusetts Study found no significant differences between FFDM and SFM in cancer detection but with significantly reduced recall rates for women imaged with FFDM. A prospective trials by Skaane et al. concurred with Lewin et al. for results in cancer detection but found higher recall rates for FFDM (Oslo I, 4.6%; Oslo II, 4.2%) when compared to SFM (Oslo I, 3.5%; Oslo II, 2.5%)… Despite these inconclusive findings, other studies have not replicated such a vast different between SFM and FFDM…Results from the Digital Mammographic Imaging Screening Trial (DMIST) report there were no differences between SFM and FFDM for the entire population, with the CDR of 0.4% and 0.44% for SFM and FFDM respectively and the recall rate was exactly the same at 8.6% for SFM and FFDM.” **Conclusion** • “Trials involving comparisons with FFDM and SFM in a screening context have demonstrated conflicting results with regards to recall rates…”
Le, M. T., Mothersill, C. E., Seymour, C. B., &	• Screen-film mammography	• N/A	• Note: the focus of this review is the rate of false positive (FP) results

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
McNeill, F. E. (2016). Is the false-positive rate in mammography in North America too high? [Review]. British Journal of Radiology, 89(1065), 20160045. **[Narrative review]**	(SFM) vs. digital mammography on false-positive rates		- Studies assessing the transition from screen-film mammography (SFM) to digital mammography in females older than 50 years have reported an increased sensitivity for detecting invasive carcinomas by digital technology…. However, the US/Canadian Digital Mammographic Imaging Screening Trial extended their study population to include females under 50 years of age (40–49 years) and subsequently reported that digital mammography did not confer significantly better diagnostic accuracy than SFM among the entire study population (40–69 years old)… However, the study did conclude that females in the 40–49-year age group, particularly those who were pre-menopausal or perimenopausal, were conferred significant benefits in terms of diagnostic accuracy from digital mammography compared with SFM. In this under 50 age group, a reduction in the FP [false positive] rate at a given diagnostic sensitivity level was also observed as a result of the shift from SFM to digital…" - "In contrast to the previously described study, other studies exist which have reported an associated increase in the FP rate following a transition to digital mammography compared with the rates previously observed during the clinical employment of SFM… Based upon modelling of a transition from film to all- digital screening in the USA, Stout et al… estimate that digital screening contributes an additional 220 FPs per 1000 females above the FP incidence seen with the current mixed use of film and digital. It can be suggested that the rise in the FP rate following the transition to digital technology is actually associated with the use of computer-aided detection (CAD) image interpretation software rather than being attributed to factors inherent to the acquisition of images by digital mammographic units themselves."
Computer-Aided Detection Systems			
Irwig L, Houssami N, van Vliet C. New technologies in screening for breast cancer: a systematic review of their accuracy. Br J Cancer. 2004;90(11):2118-2122 **[Systematic Review - Critically Low Quality]**	- Single reading with CAD vs. single reading (2 studies) - Single reading with CAD vs. double reading	- "MEDLINE was searched from 1966 to December 2002" - "The search was extended by examining references given in relevant primary studies and review articles, contact with content experts, and targeted further MEDLINE searches,	**Results** - Two studies of single reading + CAD vs. single reading report on false positive rate (FPR) based on recall rate (see table 3 of the publication): - single reader 8.5%; single reader + CAD 7.6% - single reader 6.5%; single reader + CAD 7.7% - One study of single reading + CAD vs. double reading report on incremental false positives (difficult to quantify according to the reviewers) **Conclusion** - "All of the studies examined the incremental value of CAD and showed improved sensitivity; the evidence on specificity is conflicting. It is not clear to

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
		for example on authors of earlier studies." • N = 4 articles describing 3 studies on CAD	what extent the improvement compares to other manoeuvres, such as having a second film reader."
Elmore, J. G., Armstrong, K., Lehman, C. D., & Fletcher, S. W. (2005). Screening for breast cancer. JAMA, 293(10), 1245-1256. **[Systematic Review - Critically Low Quality]**	• Single reading with CAD vs. single reading	• "searches of MEDLINE, The Cochrane Library, the National Guideline Clearinghouse Web site, the US Preventive Services Task Force recommendations and reviews,[5,13] and the International Agency for Research on Cancer Handbook of Cancer Prevention (IARC)[4] were performed to identify English-language articles about breast cancer screening" • "The bibliographies of retrieved articles were also scanned to retrieve additional relevant articles." • No CAD vs. CAD (N = 2 studies). Publication dates range from 2001 - 2004	**Results** (see table 2 of the publication) • Cancer detection rates Study 1: no CAD 3.2 per 1000; CAD 3.8 per 1000 Study 2: no CAD 3.49 per 1000; CAD 3.55 per 1000 • Recall rates Study 1: no CAD 6.5%; CAD 7.7% Study 2: no CAD 11.39%; CAD 11.4% **Conclusions** • "One study suggested that computer-aided detection increases cancer detection rates and recall rates while a second larger study did not find any significant differences." • "…presently data are limited"
Taylor, P., & Potts, H. W. (2008). Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer,	• Single reading with CAD vs. single reading	• "The NLH PubMed database was searched…. Google Scholar, Biotech, CINAHL, Embase, HMIC, Pyschinfo, Web of Science and Science Direct were searched…. The online catalogue of the British Library and recent proceedings of relevant conferences were searched. A previous systematic	**Results** • "None of the studies shows a statistically significant increase in cancer detection rate and neither group shows a pooled effect. The overall estimate of the effect is an odds ratio of 1.04 (95% confidence interval (CI): 0.96, 1.13) that is not significant (v2(1) = 0.86, p = 0.35)." "There is no evidence of heterogeneity between or within the matched and unmatched studies…" • "The evidence on the impact of CAD on recall rate…is less clear. All the studies showed increased recall rates, but there is a strong evidence of heterogeneity: overall test, χ2 (9) = 148.1, p < 0.001, I2 = 94%. The matched studies do not show heterogeneity: χ2 (4) = 3.6, p = 0.47, I2 < 0.1%."

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
44(6), 798-807. **[Systematic Review – Critically Low Quality]**		review of double reading was identified and its references were checked,3 as were references in retrieved papers." • Single reading with CAD vs. single reading (N = 10 studies). Publication dates range from 2001 – 2008	• "The overall pooled estimate for the odds ratio is 1.10 (95% CI: 1.09, 1.12), which is significant ($\chi2(1) = 130.3$, $p < 0.001$), as are the estimates for the matched and unmatched studies separately." **Conclusion** • "CAD does not have a significant effect on cancer detection rate… and increases recall rate…However, there is considerable heterogeneity in the impact on recall rate in both sets of studies."
Noble, M., Bruening, W., Uhl, S., & Schoelles, K. (2009). Computer-aided detection mammography for breast cancer screening: systematic review and meta-analysis. Archives of Gynecology & Obstetrics, 279(6), 881-890. [**Systematic Review - Moderate Qualit**y]	• CAD; unclear comparators (single or double reading) • Results on cancer detection and recall rates are based on 5 studies of single-reading with CAD vs. single reading	• "searched seventeen databases including MEDLINE, EMBASE, and the Cochrane Library though September 25, 2008, and we hand-searched the bibliographies/reference lists from peer-reviewed and gray literature (i.e. reports and studies produced by local government agencies, private organizations, educational facilities, and corporations) to identify clinical studies not identified by the electronic searches." • N = 5 studies on CAD	**Results** [focus on false-positive rates and not recall rates] • "The incremental cancer detection rate among women screened using CAD over single-read mammography alone was 50 (95% CI 30–80) women per 100,000 screened. The data from the five studies in this evidence base were not substantially heterogenous (I^2<0.001%)." • "Incremental recall of healthy women. The REMA yielded a rate of 1,190 (95% CI 1,090–1,290) additional healthy women per 100,000 screened with CAD who would not have been recalled if screened by single-read mammography alone. However, due to the unexplained heterogeneity and lack of robustness, the point estimate may not reliably indicate the incremental recall rate(s) of healthy women." • "Proportion of recalled women who were healthy. Ninety six percent (95% CI 93.9–97.3%) of women recalled based on CAD findings did not have cancer. The findings from the evidence base were not substantially heterogenous (I2<0.001%)" **Conclusion** • "We agree that CAD increases the recall of healthy women…"
Azavedo E, Zackrisson S, Mejàre I, Heibert Arnlind M. Is single reading with computer-aided detection (CAD) as good as double reading in mammography screening? A systematic	• Single reading with CAD vs. double reading	• The electronic literature search included the databases PubMed, EMBASE, and The Cochrane Library from 1950 to November 2011. All Western European languages were accepted."	**Results** • Of the four included studies, three had methodological limitations and only one study of moderate quality was included in the GRADE synthesis. This prospective multicentre study based on the UK national screening program found no significant differences between single reading with CAD and double reading for cancer detection rate (7.02 vs. 7.06 per 1000) but single reading with CAD resulted in a significantly higher recall rate compared to double reading (3.9% vs. 3.4%, P=0.001)

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
review. BMC Med Imaging. 2012 Jul 24;12:22. **[Systematic Review – Moderate Quality]**		• "Hand search and grey literature did not result in any additional articles." • Single reading + CAD vs. double reading (N = 4)	• Two of the three low-quality studies were conducted in the USA; these studies show no significant differences in cancer detection rates between single reading with CAD and double reading; results for recall rates were inconsistent. • The third low-quality study was conducted in the UK. Due to lack of follow-up, the authors calculated a relative sensitivity which was non-significantly lower with single reading + CAD (91.5%) than with double reading (98.4%). The recall rate was significantly higher with single reading + CAD: 6.1% vs. 5.0% with double reading. **Conclusion** • "The scientific evidence is insufficient to determine whether the accuracy of single reading + CAD is at least equivalent to that obtained in standard practice, i.e. double reading where two breast radiologists independently read the mammographic images."
Astley SM, Gilbert FJ. Computer-aided detection in mammography. Clin Radiol. 2004;59(5):390-399 **[Narrative review]**	• CAD	• N/A	• "Although the sensitivity of detection algorithms is approaching that of human readers for some signs of abnormality, the specificity is still relatively poor…. a system based on current computer-based methods operating at acceptable sensitivities would have a recall rate several times higher than that of an expert radiologist." • "It is likely that there will be learning effects, so ideally the readers should be given time to become familiar with the CAD system and reading with prompts before any evaluation takes place" • "If CAD is to be introduced into the programme, it will be necessary to determine what difference the system would make to a reader, whether any difference is the same for all types and levels of experience of readers, and the magnitude of that difference."
Houssami, N., Given-Wilson, R., & Ciatto, S. (2009). Early detection of breast cancer: overview of the evidence on computer-aided detection in mammography screening. Journal of Medical Imaging &	• CAD	• N/A	• "Studies show that CAD can improve the sensitivity of a single reader, with an incremental cancer detection rate (from adding CAD to a single read) ranging between 1 and 19%. However, CAD will also substantially increase the recall rate (decrease the reader's specificity) causing additional recall in approximately 6–35% of women. Evidence indicates that CAD does not perform as well as double (human) reading in the context of organized breast screening where double reading is the standard of care."

<table>
<tr><th>Reference
[AMSTAR score for systematic reviews]</th><th>Factor</th><th>Materials and Methods</th><th>Results and Authors' conclusions</th></tr>
<tr><td>Radiation Oncology, 53(2), 171-176.
[<b>Narrative review</b>]</td><td></td><td></td><td></td></tr>
<tr><td colspan="4"><b>Tomosynthesis</b></td></tr>
<tr><td>Svahn, T. M., Macaskill, P., & Houssami, N. (2015). Radiologists' interpretive efficiency and variability in true- and false-positive detection when screen-reading with tomosynthesis (3D-mammography) relative to standard mammography in population screening. Breast, 24(6), 687-693.
[<b>Systematic Review - Critically Low Quality</b>]</td><td>• Digital breast tomosynthesis as adjunct to full-field digital mammography (2D/3D) relative to 2D alone</td><td>• To identify all potentially eligible primary studies we systematically searched the literature; replicating a previously published systematic search [12] in January 2015, and further updated the search at week 1 July 2015. The search strategy consisted of a Medline search …and also contact with content experts."
• 2D/3D vs. 2D (N = 3 studies / 4 publications)</td><td><b>Results</b>
• Cancer detection rate for 2D/3D per 1000 women (reader-averaged increase relative to that of 2D; %)
STORM study: 8.1 (+53)
OTST study: 8.0 (+31)
Houston study: 5.4 (+35)
• Recall rate or False Positive rate (OTST) (reader-averaged decrease relative to that of 2D; %)
STORM study: 3.5 (-20)
OTST study: 5.3 (-13)
Houston study: 5.5 (-37.5)
<b>Conclusion</b>
• "…the majority of radiologists were more efficient screen-readers using 2D/3D-mammography (they had less FPs for each detected breast cancer) than using 2D-mammography."</td></tr>
<tr><td>Hodgson, R., Heywang-Kobrunner, S. H., Harvey, S. C., Edwards, M., Shaikh, J., Arber, M., et al. (2016). Systematic review of 3D mammography for breast cancer screening. [Review]. Breast, 27, 52-</td><td>• Digital breast tomosynthesis (DBT) (alone or with full field digital mammography (FFDM) compared with FFDM alone</td><td>• "This systematic review was carried out according to the systematic review guidance provided in the Cochrane Handbooks …The searches were performed and concluded in October 2014."
• "Reference lists of relevant papers retrieved by the searches were scanned for potentially eligible studies. Systematic reviews</td><td><b>Results</b>
• Recall rate<table><tr><th>Study</th><th>DBT=FFDM</th><th>FFDM</th></tr><tr><td colspan="3">European studies</td></tr><tr><td>STORM</td><td>4.3%</td><td>5.0%</td></tr><tr><td>OTST single reading</td><td>2.78%</td><td>2.1%</td></tr><tr><td>OTST double reading</td><td>3.67%</td><td>2.9%</td></tr></table></td></tr>
</table>

<table>
<tr><th>Reference
[AMSTAR score for systematic reviews]</th><th>Factor</th><th>Materials and Methods</th><th>Results and Authors' conclusions</th></tr>
<tr>
<td>61. <b>[Systematic Review – Moderate Quality</b>]</td>
<td></td>
<td>identified by the searches were checked for additional reported research not retrieved by the database searches. Citation searches were carried out on identified records.”
• FFDM vs. DBT+FFDM (N = 5 studies / 16 reports)</td>
<td>
<table>
<tr><td colspan="3">US studies</td></tr>
<tr><td>Destounis 2014</td><td>4.20%</td><td>11.45%</td></tr>
<tr><td>Lourenco 2014</td><td>6.40%</td><td>9.3%</td></tr>
<tr><td>Friedewald 2014</td><td>8.95%</td><td>10.57%</td></tr>
</table>

• Cancer detection rates
<table>
<tr><td>Study</td><td>DBT=FFDM</td><td>FFDM</td></tr>
<tr><td colspan="3">European studies</td></tr>
<tr><td>STORM</td><td>0.81%</td><td>0.53%</td></tr>
<tr><td>OTST single reading</td><td>0.80%</td><td>0.61%</td></tr>
<tr><td>OTST double reading</td><td>0.94%</td><td>0.71%</td></tr>
<tr><td colspan="3">US studies</td></tr>
<tr><td>Destounis 2014</td><td>0.57%</td><td>0.38%</td></tr>
<tr><td>Lourenco 2014</td><td>0.46%</td><td>0.54%</td></tr>
<tr><td>Friedewald 2014</td><td>0.55%</td><td>0.43%</td></tr>
</table>

<b>Conclusions</b>

• “Evidence suggests that recall and false positive rates may be lower using DBT + FFDM, especially for single reader paradigms such as those common in the US.”

• “Overall, the evidence suggests that cancer detection rates and invasive cancer detection rates are higher using DBT + FFDM than with FFDM, but non-invasive cancer detection rates are unchanged.”
</td>
</tr>
</table>

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
			• Note: no data reported on DBT alone
Nelson, H. D., Pappas, M., Cantor, A., Griffin, J., Daeges, M., & Humphrey, L. (2016). Harms of Breast Cancer Screening: Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation. Annals of Internal Medicine, 164(4), 256-267. **[Systematic Review – Moderate Quality]**	• "mammography and tomosynthesis versus mammography alone" • Note: the scope of this review is broad and includes multiple screening modalities and outcomes	• "A research librarian conducted electronic searches of the Cochrane Central Register of Controlled Trials, the Cochrane Database of Systematic Reviews, and Ovid MEDLINE through December 2014 for relevant studies and systematic reviews. Searches were supplemented by references identified from additional sources, including reference lists and experts. Studies of harms included in the previous systematic review for the USPSTF (2, 3) were also included." • DM vs. DM+tomosynthesis (N = 5 Studies)	• "Four of 5 studies showed statistically significantly lower rates of recall for tomosynthesis and mammography than for mammography alone."
Pozz, A., Corte, A. D., Lakis, M. A., & Jeong, H. (2016). Digital Breast Tomosynthesis in Addition to Conventional 2D Mammography Reduces Recall Rates and is CostEffective. Asian Pacific Journal of Cancer Prevention: Apjcp, 17(7), 3521-3526. [**Systematic**	• Digital breast tomosyntheis (DBT) vs. digital mammography (DM) • DBT+DM vs. DM	• "A comprehensive systematic review was conducted independently by all three authors using search terms such as tomosynthesis, breast imaging, 3D-mammography. PubMed, Medline, Google Scholar, Ovid, and Cochrane data search engines were utilized from inception until April 2016. The authors then manually scrutinized reference lists in the recovered articles and	**Results** • DBT+DM vs DM Ciatto et al. 2013 Potential 17.2% reduction in recall rate using DBT. CDR 8.1/1000 in DBT+DM vs 5.3/1000 in DM alone Conant et al. 2016 Recall rate 8.7% in DBT+DM vs 10.4% in DM CDR 5.9/1000 in DBT+DM vs 4.4/1000 in DM Destounis et al. 2014 Recall rate 4.5% in DBT+DM vs 11.45% in DM CDR 5.7/1000 in DBT+DM vs 3.8/1000 in DM

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
Review - Critically Low Quality]		relevant abstracts from scientific meetings to identify any further articles." • DBT vs. digital mammography (N=3 references) • DBT+DM vs. DM (N=10/11 references report on outcomes of interest)	Durand et al. 2015 Recall rate 7.8% in DBT+DM vs 12.3% in DM. CDR 5.9/1000 in DBT/DM vs 5.7/1000 in DM. Friedewald et al. 2014 Recall rate 9,1% in DBT+DM, vs 10,7% in DM alone. CDR 5.4/1000 in DBT+DM vs 4.2/1000 in DM alone. Gilbert et al. 2015 (reports on sensitivity and specificity) Greenberg et al. 2014 Recall rate 13,6% in DBT+DM vs 16,2% in DM. CDR 6.3/1000 in DBT+DM vs 4.9/1000 in DM. Haas et al. 2013 Recall rate 8.4% in DBT+DM vs 12% in DM alone. CDR 5.4/10000 in DBT+DM vs 4.2/1000 in DM alone. Rose et al. 2013 Recall rate 5.5% in DBT+DM vs 8.7% in DM. CDR 5.37/1000 in DBT+DM vs 4.04/1000 in DM Skaane et al. 2013 FP 53.1/1000 in DBT+DM vs 61.1/1000 in DM. CDR 8.0/1000 in DBT+DM vs 6.1/1000 in DM. PPV recall 16.2% in DBT+DM vs 6% in DM Sumkin et al., 2015 Recall rate 25.5% in DBT+DM vs 38.4%in DM. • DBT vs. DM Lang et al. 2016 Recall rate 3.8% in DBT vs 2.6% in DM. CDR 8.9/1000 in DBT vs 6.3/1000 in DM. Lourenco et al. 2015

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
			Overall recall rate 6.4% in DBT vs 9.3% in DM. No significant differences regarding biopsy PPV and CDR. McDonald et al. 2015 Recall rate 16% in DBT, 20.5% in DM at first screening; 7.8% in DBT vs 9.1% in DM for subsequent screenings. CDR 5.9/1000 in DBT, vs 4.2/1000 in DM (first screening); 5.4/1000 in DBT, vs 4.6/1000 in DM (subsequent screenings). **Conclusion** • "In conclusion, Digital breast tomosynthesis addresses the primary limitations of conventional screening mammography by increasing conspicuity of invasive cancers while concomitantly reducing false-positive results. This results in a significant reduction in recall rates."
Cole, E. B., & Pisano, E. D. (2016). Tomosynthesis for breast cancer screening. Clinical Imaging, 40(2), 283-287. [**Narrative review**]	• Hologic two-view TM plus two-view digital mammography (DM) • Hologic two-view TM plus synthetic two-dimensional (2D) mammography (sDM) • General Electric's (GE) Healthcare one-view TM (MLO) plus one-view DM (CC)	• "A search was conducted for papers published between January 1, 2013, and March 28, 2015, with the search terms [breast AND tomosynthesis [ti] AND screening [ti]]" • N=22 article. • Note: although the authors report on keywords used, numbers of articles retained and excluded (with reasons), they do not classify their review as systematic.	**Results** • "The FDA SSED [Summary of Safety and Effectiveness Data] for the SenoClaire, which is approved for screening using this one-view 3D TM MLO plus one-view 2D DM CC acquisition protocol, demonstrates a reduction in recall rate and improved specificity versus DMalone,with AUC and sensitivity basically equivalent…" • FDA summary of safety and effectiveness data SenoClaire-P130020. Recall rate (95% CI) TM MLO plus DM CC: 0.340 (0.308, 0.373) TM MLO: 0.348 ((0.316, 0.380) Two-View DM: 0.406 (0.374, 0.438) AUC (95% CI) TM MLO plus DM CC: 0.842 (0.786, 0.899) TM MLO: 0.820 (0.752, 0.888) Two-View DM: 0.853 (0.798, 0.908) **Conclusion** • "There is also strong evidence that the Hologic system (TM plus DM) results in fewer recalls from screening mammography. These results are very promising."

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
			- The authors acknowledged that there was lack of published data on the use of the Hologic system with TM plus synthetic DM or on the system manufactured by GE Healthcare.
Gilbert, F. J., Tucker, L., & Young, K. C. (2016). Digital breast tomosynthesis (DBT): a review of the evidence for use as a screening tool. Clinical Radiology, 71(2), 141-150. [**Narrative review**]	- digital breast tomosynthesis (DBT) alone and in combination with FFDM	- N/A	- "Superimposition of normal tissues may produce features on mammography, which are suspicious for cancer and lead to unnecessary recall for further assessment and diagnostic tests to exclude malignancy. By facilitating the analysis of superimposed breast structures, DBT may enable the reader to identify features that, for example, appear to be asymmetric density on FFDM image as normal composite shadows, thereby decreasing the number of false-positive recalls, … associated health costs, … and reducing patient anxiety..." - "In general, studies have demonstrated the potential for DBT to decrease recall rates and increase cancer detection rates; however, the use of DBT systems with different technical configurations coupled with variations in study methodologies and case configurations have produced conflicting results regarding the efficacy of DBT." - "Results also show cancers being detected at a smaller size and a decrease in false-positive recall rates of 15e20%."
Vedantham, S., Karellas, A., Vijayaraghavan, G. R., & Kopans, D. B. (2015). Digital Breast Tomosynthesis: State of the Art. Radiology, 277(3), 663-684. [**Narrative review**]	- digital breast tomosynthesis (DBT) alone or in combination with FFDM	- N/A	- Studies in screening populations show a statistically significant reduction in recall rate with two view DBT plus full-field digital mammography (FFDM) compared with two-view FFDM." - "Prospective trials in screening population from Europe show a statistically significant increase in cancer detection rate with two view DBT plus FFDM compared with two-view FFDM, and retrospective observational studies from the United States show either a significant or a nonsignificant increase."
Skaane, P. (2017). Breast cancer screening with digital breast tomosynthesis. Breast Cancer, 24(1), 32-41. [**Narrative review**]	- FFDM+DBT vs. FFDM alone	- N/A	- "The retrospective screening studies from USA have all shown a significant decrease in the recall rate using DBT as adjunct to mammography. Most of these studies have also shown an increase in the cancer detection rate, and the non-significant results in some studies might be explained by a lack of statistical power. All the three prospective European trials have shown a significant increase in the cancer detection rate."

Table A5. Quality assurance practices

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
Reading Volume			
Mohd Norsuddin, N., Reed, W., Mello-Thoms, C., & Lewis, S. J. (2015). Understanding recall rates in screening mammography: A conceptual framework review of the literature. Radiography, 21(4), 334-341. [**Narrative review**]	• Annual reading volume	• "The search of the literature was conducted in MEDLINE, CINAHL (EbSCOhost), SPIE library,Web of Science, PubMed, Scopus databases and Google Scholar. No specific year of publication was imposed in this search however, we prioritised studies from 2000 onwards which were likely to capture current imaging modalities in screening mammography." • Note: although the authors report on databases searched and keywords used, they do not classify their review as systematic.	• "Reading high volumes of mammograms is likely to improve the identification of the normal variations in breast tissues allowing readers to be more certain of benign findings and identify a range of cancers presentations… It has been shown that readers who read a larger number of mammograms have lower false positive results, because these readers have developed a better knowledge bank of normal presentations seen on screening mammograms…Various breast screening programmes and organizations have set minimum annual volumes of cases to be read by breast readers. This volume varies across countries; from 960 cases during a 24 months period in the United States… 2000 cases per year in Australia… and as high as 5000 mammograms per year in the UK…It is suggested that a minimum case threshold may allow readers to reduce the number of women recalled for further assessment (abnormal interpretation rate) and to increase the CDR."
Le, M. T., Mothersill, C. E., Seymour, C. B., & McNeill, F. E. (2016). Is the false-positive rate in mammography in North America too high? [Review]. British Journal of Radiology, 89(1065), 20160045. [**Narrative review**]	• Annual reading volume	• N/A	• "An alternative factor which may contribute to high recall rates is the difference in reporting experience required of radiologists interpreting mammograms in the USA and Canada compared with other jurisdictions such as the UK. While the US Mammography Quality Standards Act considers the reporting of only 480 mammograms per annum to be adequate, … the UK mandates that practising radiologists must read a minimum of 5000 mammograms per year to continue practising in mammography specialization… The drastic contrast between these requirements suggests a clear difference in the experience that UK radiologists acquire early on in their careers over US radiologists and thus potentiates a propensity to generate reports with greater certainty and accordingly fewer recalls. Similar to the US regulations, the Canadian Mammography Quality Guidelines also only require that radiologists report at least 480 mammograms per year to maintain their qualification… It can be suggested that this more lax requirement contributes to the relatively high FP and recall rates that can be observed in North

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
			America, as reading volume has been shown to have a significant impact upon the sensitivity and specificity of mammographic reporting…"
Double Reading			
Taylor, P., & Potts, H. W. (2008). Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer, 44(6), 798-807. [**Systematic Review Critically Low Quality**]	• Double reading vs. single reading	• "The NLH PubMed database was searched…. Google Scholar, Biotech, CINAHL, Embase, HMIC, Pyschinfo, Web of Science and Science Direct were searched…. The online catalogue of the British Library and recent proceedings of relevant conferences were searched. A previous systematic review of double reading was identified and its references were checked,3 as were references in retrieved papers." • The authors state that studies in which the second reader was not a radiologist were included. It is not clear whether second reader being a radiologist was a criterion for exclusion. Based on the US NLM summary[8], it was not. • Double reading vs. single reading (N = 17). Publication dates range from 1991 – 2008	**Results** • Of the 17 included studies, arbitration was used for resolution of discrepancies in 5, consensus in 3, mixed practice was adopted in 3 and unilateral recall in 6. • "There is clear evidence of heterogeneity: overall test, $\chi^2(16) = 925.7$, $p < 0.001$, $I^2 = 98\%$. There is heterogeneity between the three groups (χ^2 (2) = 513.5, $p < 0.001$) and within each of the groups (for arbitration/consensus studies, χ^2 (7) = 306.5, $p < 0.001$, $I^2 = 98\%$; for mixed studies, χ^2 (2) = 8.6, p = 0.014, $I^2 = 77\%$; for unilateral studies, χ^2 (5) =97.2, $p < 0.001$, $I^2 = 95\%$). • <u>Recall rates</u>. "All the mixed and unilateral studies show increases in recall rate. Overall, arbitration studies show a decrease, but two, including one of the largest studies… show a significant increase" • <u>Recall rates</u>. The overall pooled estimate for the odds ratio (95% CI) arbitration/consensus: 0.94 (0.92, 0.96) mixed: 1.21 (1.19, 1.24) unilateral: 1.31 (1.29, 1.33) overall: 1.17 (1.15, 1.18) • <u>Cancer detection rates</u>. There is no evidence of heterogeneity: overall test, χ^2 (16) = 5.1, p = 1.0, $I^2 < 0.1\%$; testing between the three subgroups, χ^2 (2) = 1.4, p = 0.50. Although individually the reported effects are mostly not significant, the pooled estimate is significant (95% CI: 1.06, 1.14; χ^2 (1) = 23.5, $p < 0.001$). • <u>Cancer detection rates</u>. The overall pooled estimate for the odds ratio (95% CI) arbitration/consensus: 1.08 (1.02, 1.15) mixed: 1.07 (0.99, 1.15) unilateral: 1.13 (1.06, 1.19) overall: 1.10 (1.06, 1.14) **Conclusion**

[8] See "Study selection" at: https://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0025540/

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
			• "There is evidence that double reading increases cancer detection rate and that double reading with arbitration does so whilst lowering recall rate."
Hackney, L., Szczepura, A., Moody, L., & Whiteman, B. (2017). Review of the evidence on the use of arbitration or consensus within breast screening: A systematic scoping review. Radiography (Lond), 23(2), 171-176. [**Systematic Review – Moderate Quality**]	• Effectiveness of arbitration and consensus as methods for resolution of discordant opinions • Although recall rates are reported, the focus is cancer detection	• "Literature searches of PubMed, Medline, CINAHL, EMBASE, Scopus, Web of Science and the Cochrane Library were supplemented by a broad Google scholar web search. Hand searching of key peer-reviewed breast and radiology journals, a manual search of reference lists and key author searching was undertaken. Grey literature was sourced by hand searching of conference proceedings and doctoral theses. Personal contact with experts internationally was also undertaken in locating relevant literature." • N = 26 studies	**Results** • Arbitration. Overall, studies reported that compared to highest reader recall (non-arbitration), arbitration resulted in significant reductions in recall rates, with relative decreases in the range of 17.8%...to 40.9%." • "With such variation in recall rates the PPV of assessment cases following arbitration is also unpredictable with low PPV's of 8.3%... to 31.2%... reported." • "There is disparity between the studies regarding the effect of arbitration on cancer detection rates." • Consensus. Of five studies that mention consensus, only two investigated its effectiveness. • "There was a supposition from some of the literature that fewer cancers will be missed by panel consensus compared to single reader arbitration. However, no evidence was found to support this." • "As with all group meetings, the dynamics within the consensus team can be a significant factor affecting the final decision." Examples: "one reader is the dominant and opinions are not equally weighted"; "individuals may change their judgment to what they 'believe others want to hear'" • "Within a number of studies…it is not possible to differentiate the effect of arbitration versus consensus as the processes are either integrated in the discussion, or both are undertaken within the decision making strategy i.e. mutual consensus between the two readers with persistent discordant case being reviewed by an arbitration panel." • "Definitions of consensus and arbitration are not clear-cut. The two terms are used interchangeably and often confusing with some studies reporting 'arbitration by an individual', others 'arbitration by a panel', and 'consensus based arbitration'. The lack of clear definitions makes it not only difficult to review the literature and synthesise the findings, but it also adds to confusion in a clinical setting when discussing processes with no clear delineations." **Conclusion** • "The insufficiency of follow-up or reporting of true interval cancers compromised the ability to conclude the effectiveness of the processes."
Posso, M., Puig, T., Carles, M., Rue, M.,	• Double reading vs. single reading	• "Databases were searched from 1st January 1990 to 20th February 2017,	**Results**

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
Canelo-Aybar, C., & Bonfill, X. (2017). Effectiveness and cost-effectiveness of double reading in digital mammography screening: A systematic review and meta-analysis. European Journal of Radiology, 96, 40-49. [**Systematic Review – Moderate Quality**]	of digital mammograms - Focus is on false positives instead of on recall rate	including Medline, EMBASE, and the Cochrane Library" - "We included studies we deemed as relevant based on our previous experience, and hand searched the bibliography of the included studies" - Double reading vs. single reading of digital mammograms (N= 2 studies reporting on false positives; N=3 studies reporting on cancer detection rate)	- The pooled proportion of false-positives results of double reading was 47.03 per 1000 screens (CI: 39.13–55.62 per 1000) and it was 40.60 per 1000 (CI: 38.58–42.67 per 1000) for single reading (P = 0.12) - The pooled cancer detection rate of double reading was 6.01 per 1000 screens (CI: 4.47–7.77 per 1000), that was not statistically different from 5.65 per 1000 screens (CI: 3.95–7.65 per 1000) observed in single reading (P= 0.76) **Conclusion** - "…too little is currently known about the effectiveness and cost-effectiveness of double reading compared with single reading in the context of digital mammography. The uncertainty on the immediate effects such as cancer detection and false positives of double reading remains large and there are no publications on long-term health outcomes. Double reading seems to increase operational costs, have a not significantly higher false-positive rate, and a similar cancer detection rate."
Mohd Norsuddin, N., Reed, W., Mello-Thoms, C., & Lewis, S. J. (2015). Understanding recall rates in screening mammography: A conceptual framework review of the literature. Radiography, 21(4), 334-341. [**Narrative Review**]	- Double reading vs. single reading	- "The search of the literature was conducted in MEDLINE, CINAHL (EbSCOhost), SPIE library,Web of Science, PubMed, Scopus databases and Google Scholar. No specific year of publication was imposed in this search however, we prioritised studies from 2000 onwards which were likely to capture current imaging modalities in screening mammography." - Note: although the authors report on databases searched and keywords used, they do not classify their review as systematic.	- "The probability a woman may be recalled has been found to be higher if only one reader considers the mammogram abnormal and this can contribute to the higher percentage of recall rates among screened women in USA population…" - "When compared with single-reading, double interpretation of screening mammograms has been shown to improve CDR… especially for less experienced readers…" - "In a large UK study, researchers postulated that arbitration and consensus between readers using the double reader strategy can lower recall rates, especially in detecting high difficulty cancers…Furthermore, they suggested that the double reading of screening mammograms can increase the cancer detection rates when compared to single reading without an overall increase in the recall rates across the screened population…"
Le, M. T., Mothersill, C. E., Seymour, C. B., &	-	-	- "Double reading (i.e. two radiologists reading each mammogram) can change the FP rate depending on the manner in which it is conducted. Independent

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
McNeill, F. E. (2016). Is the false-positive rate in mammography in North America too high? [Review]. British Journal of Radiology, 89(1065), 20160045. [**Narrative Review**]			radiologist reporting has shown an increase in the probability for FPs because patients are recalled if either one of the radiologists considers the mammogram to be abnormal… On the other hand, arbitration and consensus between the two readers has resulted in a significant decrease in the FP rate…It has also been reported that blinded double reading, whereby the second reader was not aware of the first reader's suggestions, significantly increased the FP rates compared with a situation where double reading was not blinded… The advantage of blind double reading, however, is the significant improvement in cancer detection sensitivity associated…"
Audit/Performance Feedback			
Soh, B. P., Lee, W., Kench, P. L., Reed, W. M., McEntee, M. F., Poulos, A., et al. (2012). Assessing reader performance in radiology, an imperfect science: lessons from breast screening. Clinical Radiology, 67(7), 623-628 [**Narrative Review**]	•	• N/A	• "it may take 2 years for falling performance to be identified by clinical audit, then another 2 years to demonstrate improvement in performance following introduction of a quality improvement programme. For some breast screening centres (or individual breast screen readers) that screen a comparatively low volume of women each year, the period of time taken to identify underperformance may be lengthened even further. A study that investigated the relationship between the UK National Health Service (NHS) breast screening programme performance and size (number of women screened in a year) of individual programme found that it was impossible to tell if small programmes were underperforming even after consolidating all data collected in 3 years. Due to statistical instability from relatively small number of data, it was suggested that low volume programmes would most likely be overlooked by clinical audits even if they were to underperform. The same problem is also applicable to individual mammography screen readers. Relying solely upon clinical audit to identify underperformance may result in a prolonged period of time during which underperformance has the potential to harm women participating in a breast screening programme. It is clear that another more efficient method of measuring readers' performance, such as screen reader test sets is needed, especially for low-volume screening centres or individual breast screen readers" • Examples of standardized mammographic screen reading test sets: the Personal Performance in Mammographic Screening (PERFORMS) test implemented by the National Health Service Breast Screening Programme (NHSBSP) in the UK in 1991; BREAST (Breastscreen Reader Assessment STrategy) introduced in Australia in 2011

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
			• Benefits of standardized test sets: "ease of application, immediacy of results, and quicker assessment of quality improvement plans". • However, "…these test case results must be validated against real clinical reading performance." • "standardized screening test sets… suffer from experimental confounders, thus questioning the relevance of these laboratory-type screening test sets to clinical performance." • "Four key factors that impact on the external validity of screening test sets were identified: the nature and extent of scrutiny of one's action, the artificiality of the environment, the oversimplification of responses, and prevalence of abnormality." • "there is little evidence demonstrating that performance in tests are strongly correlated to actual performance in the clinical setting."
Comparison with Prior Mammograms			
None identified			
Number of Mammographic Views			
Le, M. T., Mothersill, C. E., Seymour, C. B., & McNeill, F. E. (2016). Is the false-positive rate in mammography in North America too high? [Review]. British Journal of Radiology, 89(1065), 20160045. [**Narrative Review**]	• two views (craniocaudal and mediolateral oblique) vs. a single view (mediolateral oblique)	• N/A	• "it has been shown that the acquisition of two views (craniocaudal and mediolateral oblique) decreases the FP rate significantly… owing to the availability of additional information to the radiologist…" • "Since the 1980s, standard practice in both the USA and Canada sees that screening mammography is conducted using two views at first and successive screening rounds…From this observation, it can be suggested that the long-practised acquisition of two views during mammography screening in North America has not contributed to the FP rates observed; rather this practice is expected to have assuaged the incidence of FPs. It is therefore apparent that other factors, and not the number of views taken during mammographic screening in North America, contribute to the FP incidence that currently exists."
Mammographic Compression			
None identified			

Table A6. Radiologist characteristics

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors' conclusions
Training, Education, and Experience			
van den Biggelaar, F. J., Nelemans, P. J., & Flobbe, K. (2008). Performance of radiographers in mammogram interpretation: a systematic review. Breast, 17(1), 85-90. [**Systematic Review - Critically Low Quality**]	- performance of radiographers (also referring to technologists and physician assistants) vs. performance of radiologists; the effect of training programmes offered to mammogram readers - performance metrics: sensitivity, specificity, PPV, NPV and diagnostic odds ratio (DOR)[9]	- PUBMED and EMBASE databases were searched up to December 2006; no limit on publication date was used. - N=6 eligible studies (N=5 reporting on the comparability of sensitivity and specificity between radiologists and radiographers; N=3 studies reporting on performance before and after the training period)	**Results** - Performance of radiographers Four of five studies demonstrated lower specificity of radiographers (64% to 91%) as compared to radiologists (81% to 95%) and comparable sensitivity of the two groups (73% to 90% for radiographers vs. 73% to 86% for radiologists). The DORs were lower for radiographers compared with radiologists One study reported a higher sensitivity for radiographers compared with radiologists - Effects of training All three studies demonstrated an increased DOR. In two studies the DOR was increased mainly due to increased specificity; in the third study, the DOR was increased due to an increased sensitivity despite a decrease in specificity. **Conclusion** - "The results showed that radiographers scored higher false positive rates with a similar sensitivity in the detection of malignancies, compared with radiologists. Furthermore, it was indicated that training programmes could improve the performance by reducing the number of false positive results and increasing the specificity."
Mohd Norsuddin, N., Reed, W., Mello-Thoms, C., & Lewis, S. J. (2015). Understanding recall rates in screening mammography: A conceptual framework	- Fellowship training; years of experience	- N/A	- "Dedicated training and specialized education in breast imaging has been shown to produce better observer performance, especially in reducing the number of recalled women alongside higher cancer detection rates.". - The authors of this review refer to two studies by a similar authorship team (Miglioretti et al. 2009 and Elmore et al. 2009) that address the benefits of fellowship training. Miglioretti et al.2009 demonstrated that radiologists who undertook a dedicated fellowship training in breast imaging had reduced false positive recall rates and met an acceptable national performance standard earlier in their career as compared to radiologists who learned through years

[9] "A DOR of 1 implies that the test has no discriminatory power at all; the larger the DOR, the better the test discriminates between patients with and without the disease of interest."

Reference [AMSTAR score for systematic reviews]	Factor	Materials and Methods	Results and Authors’ conclusions
review of the literature. Radiography, 21(4), 334-341. [**Narrative Review**]			of experience. The authors of both articles report that only a small proportion (<10%) of radiologists who read screening mammograms are fellowship-trained. • “Readers with more years of experience read and interpret images faster and can identify lesions more accurately… and eye tracking analysis has also shown that incorrect decisions such as false positives and false negatives actually attract extended visual time, with correctly identified cancers visually located more quickly and efficiently.”
Age and Gender			
None Identified			
Litigation Concerns			
Le, M. T., Mothersill, C. E., Seymour, C. B., & McNeill, F. E. (2016). Is the false-positive rate in mammography in North America too high? [Review]. British Journal of Radiology, 89(1065), 20160045. [**Narrative Review**]	•	•	• “…the perceived risk of severe litigation consequences associated with medical malpractice in the USA has been identified as a possible contributor to the relatively high recall rates that exist in the USA…” • “Elmore et al…surveyed 124 US radiologists in an effort to assess the relationship between radiologists experience with diagnosis specific malpractice in the mammography setting and their recall rates subsequent to those experiences. Prior involvement in a mammography-related medical malpractice case did not increase the recall rate or FP rate above that observed in those who were not involved in prior litigation claims. Further evidence suggesting that radiologist perception is not always predictive of his/her reporting patterns can be observed in the results published by a study investigating the effect of introducing new breast density reporting laws upon radiologist reporting.” • “Despite these results, litigation risk should not be discounted as a potential contributing factor to the heightened recall rates observed in North America when compared with other jurisdictions, since the volume of malpractice claims put forth in the USA is actually substantially greater than that in European countries such as Italy and the Netherlands.” • “Because a greater risk of litigation is a reality in the USA, it is possible that it has led to an overall greater awareness of litigation risk among all North American radiologists. This in turn could contribute to an overall greater recall and FP rate compared with other jurisdictions where this pressure does not weigh upon practice as heavily.”

Appendix 4. List of original studies

Table A7. Technology

#	Reference	Comments
	Screen-film vs. Digital Mammography [Search Start Date: 2009]	
1.	Arrospide, A., Comas, M., Mar, J., Sala, M., Hernandez, C., Roman, R., et al. (2011). Budget impact analysis of switching to digital mammography in a breast cancer population-based screening program. Value in Health, 14 (7), A438.	Conference abstract Full text article: Comas et al. 2014[10]
2.	Campari, C., Giorgi Rossi, P., Mori, C. A., Ravaioli, S., Nitrosi, A., Vacondio, R., et al. (2016). Impact of the Introduction of Digital Mammography in an Organized Screening Program on the Recall and Detection Rate. Journal of Digital Imaging, 29(2), 235-242.	
3.	Chiarelli, A. M., Edwards, S. A., Prummel, M. V., Muradali, D., Majpruz, V., Done, S. J., et al. (2013). Digital compared with screen-film mammography: performance measures in concurrent cohorts within an organized breast screening program. Radiology, 268(3), 684-693.	
4.	de Munck, L., de Bock, G. H., Otter, R., Reiding, D., Broeders, M. J., Willemse, P. H., & Siesling, S. (2016). Digital vs screen-film mammography in population-based breast cancer screening: performance indicators and tumour characteristics of screen-detected and interval cancers. British Journal of Cancer, 115(5), 517-524.	
5.	Dabbous, F., Dolecek, T. A., Friedewald, S. M., Tossas-Milligan, K. Y., Macarol, T., Summerfelt, W. T., et al. (2017). Performance characteristics of digital vs film screen mammography in community practice. Breast J. Breast J. 2017 Nov 5. doi: 10.1111/tbj.12942. [Epub ahead of print]	Added from other source
6.	Feeley, L., Kiernan, D., Mooney, T., Flanagan, F., Hargaden, G., Kell, M., et al. (2011). Digital mammography in a screening programme and its implications for pathology: a comparative study. [Comparative Study]. Journal of Clinical Pathology, 64(3), 215-219	

[10] https://www.ncbi.nlm.nih.gov/pubmed/24832200

#	Reference	Comments
7.	Glynn, C. G., Farria, D. M., Monsees, B. S., Salcman, J. T., Wiele, K. N., & Hildebolt, C. F. (2011). Effect of transition to digital mammography on clinical outcomes. Radiology, 260(3), 664-670.	
8.	Hambly, N. M., McNicholas, M. M., Phelan, N., Hargaden, G. C., O'Doherty, A., & Flanagan, F. L. (2009). Comparison of digital mammography and screen-film mammography in breast cancer screening: a review in the Irish breast screening program. AJR. American Journal of Roentgenology, 193(4), 1010-1018.	
9.	Hofvind, S., Skaane, P., Elmore, J. G., Sebuodegard, S., Hoff, S. R., & Lee, C. I. (2014). Mammographic performance in a population-based screening program: before, during, and after the transition from screen-film to full-field digital mammography. [Comparative Study]. Radiology, 272(1), 52-62.	
10.	Juel, I. M., Skaane, P., Hoff, S. R., Johannessen, G., & Hofvind, S. (2010). Screen-film mammography versus full-field digital mammography in a population-based screening program: The Sogn and Fjordane study. Acta Radiologica, 51(9), 962-968.	
11.	Karssemeijer, N., Bluekens, A. M., Beijerinck, D., Deurenberg, J. J., Beekman, M., Visser, R., et al. (2009). Breast cancer screening results 5 years after introduction of digital mammography in a population-based screening program. Radiology, 253(2), 353-358.	
12.	Lipasti, S., Anttila, A., & Pamilo, M. (2010). Mammographic findings of women recalled for diagnostic work-up in digital versus screen-film mammography in a population-based screening program. Acta Radiologica, 51(5), 491-497	
13.	Perry, N. M., Patani, N., Milner, S. E., Pinker, K., Mokbel, K., Allgood, P. C., et al. (2011). The impact of digital mammography on screening a young cohort of women for breast cancer in an urban specialist breast unit. European Radiology, 21(4), 676-682.	
14.	Sala, M., Comas, M., Macia, F., Martinez, J., Casamitjana, M., & Castells, X. (2009). Implementation of digital mammography in a population-based breast cancer screening program: effect of screening round on recall rate and cancer detection. Radiology, 252(1), 31-39.	
15.	Sala, M., Salas, D., Belvis, F., Sanchez, M., Ferrer, J., Ibanez, J., et al. (2011). Reduction in false-positive results after introduction of digital mammography: analysis from four population-based breast cancer screening programs in Spain. Radiology, 258(2), 388-395.	

16.	Sala, M., Domingo, L., Macia, F., Comas, M., Buron, A., & Castells, X. (2015). Does digital mammography suppose an advance in early diagnosis? Trends in performance indicators 6 years after digitalization. European Radiology, 25(3), 850-859.	
17.	Sankatsing, V. D. V., Fracheboud, J., de Munck, L., Broeders, M. J. M., van Ravesteyn, N. T., Heijnsdijk, E. A. M., et al. (2018). Detection and interval cancer rates during the transition from screen-film to digital mammography in population-based screening. BMC Cancer, 18(1), 256.	
18.	Theberge, I., Vandal, N., Langlois, A., Pelletier, E., & Brisson, J. (2016). Detection Rate, Recall Rate, and Positive Predictive Value of Digital Compared to Screen-Film Mammography in the Quebec Population-Based Breast Cancer Screening Program. [Comparative Study]. Canadian Association of Radiologists Journal, 67(4), 330-338.	
19.	van Luijt, P. A., Fracheboud, J., Heijnsdijk, E. A., den Heeten, G. J., de Koning, H. J., & National Evaluation Team for Breast Cancer Screening in Netherlands Study, G. (2013). Nation-wide data on screening performance during the transition to digital mammography: observations in 6 million screens. European Journal of Cancer, 49(16), 3517-3525.	
20.	Van Ongeval, C., Van Steen, A., Vande Putte, G., Zanca, F., Bosmans, H., Marchal, G., et al. (2010). Does digital mammography in a decentralized breast cancer screening program lead to screening performance parameters comparable with film-screen mammography? European Radiology, 20(10), 2307-2314.	
21.	van Ravesteyn, N. T., Miglioretti, D. L., Stout, N. K., Lee, S. J., Schechter, C. B., Buist, D. S., et al. (2012). Tipping the balance of benefits and harms to favor screening mammography starting at age 40 years: a comparative modeling study of risk. [Research Support, N.I.H., Extramural]. Annals of Internal Medicine, 156(9), 609-617.	
22.	Vernacchia, F. S., & Pena, Z. G. (2009). Digital mammography: its impact on recall rates and cancer detection rates in a small community-based radiology practice. AJR. American Journal of Roentgenology, 193(2), 582-585.	
23.	Vinnicombe, S., Pinto Pereira, S. M., McCormack, V. A., Shiel, S., Perry, N., & Dos Santos Silva, I. M. (2009). Full-field digital versus screen-film mammography: comparison within the UK breast screening program and systematic review of published data. Radiology, 251(2), 347-358.	
	-	

	Single Reading + CAD vs. Single Reading [Search Start Date: 2008]	
2.	Fenton, J. J., Abraham, L., Taplin, S. H., Geller, B. M., Carney, P. A., D'Orsi, C., et al. (2011). Effectiveness of computer-aided detection in community mammography practice. J Natl Cancer Inst, 103(15), 1152-1161.	Added from other source
3.	Gromet, M. (2008). Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. AJR. American Journal of Roentgenology, 190(4), 854-859.	
4.	Lehman, C. D., Wellman, R. D., Buist, D. S., Kerlikowske, K., Tosteson, A. N., Miglioretti, D. L., et al. (2015). Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med, 175(11), 1828-1837.	Added from other source
5.	Sanchez Gomez, S., Torres Tabanera, M., Vega Bolivar, A., Sainz Miranda, M., Baroja Mazo, A., Ruiz Diaz, M., et al. (2011). Impact of a CAD system in a screen-film mammography screening program: a prospective study. European Journal of Radiology, 80(3), e317-321.	
1.	Bargallo, X., Santamaria, G., Del Amo, M., Arguis, P., Rios, J., Grau, J., et al. (2014). Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. European Journal of Radiology, 83(11), 2019-2023.	
2.	Gromet, M. (2008). Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. AJR. American Journal of Roentgenology, 190(4), 854-859.	
	CAD vs. a Third Reader as an Arbitrator in a Double-Reading program	
1.	James, J. J., & Cornford, E. J. (2009). Does computer-aided detection have a role in the arbitration of discordant double-reading opinions in a breast-screening programme? Clinical Radiology, 64(1), 46-51.	
	Tomosynthesis [Search Start Date: 2014]	

#	Reference	Comments
1.	Friedewald, S. M., Rafferty, E. A., Rose, S. L., Durand, M. A., Plecha, D. M., Greenberg, J. S., et al. (2014). Breast cancer screening using tomosynthesis in combination with digital mammography. JAMA, 311(24), 2499-2507.	
2.	Giess, C. S., Pourjabbar, S., Ip, I. K., Lacson, R., Alper, E., & Khorasani, R. (2017). Comparing Diagnostic Performance of Digital Breast Tomosynthesis and Full-Field Digital Mammography in a Hybrid Screening Environment. AJR Am J Roentgenol, 209(4), 929-934.	Added from other source
3.	Greenberg, J. S., Javitt, M. C., Katzen, J., Michael, S., & Holland, A. E. (2014). Clinical performance metrics of 3D digital breast tomosynthesis compared with 2D digital mammography for breast cancer screening in community practice. AJR. American Journal of Roentgenology, 203(3), 687-693.	
4.	Hogue, J. C., Julien, M., Loisel, Y., Provencher, L., & Diorio, C. (2016). Improved detection rate of invasive breast cancers with tomosynthesis compared to 2D mammography in a screening program context. European Journal of Cancer, 2), S148.	Conference abstract; PDF not found
5.	Houssami, N., Bernardi, D., Pellegrini, M., Valentini, M., Fanto, C., Ostillio, L., et al. (2017). Breast cancer detection using single-reading of breast tomosynthesis (3D-mammography) compared to double-reading of 2D-mammography: Evidence from a population-based trial. Cancer Epidemiology, 47, 94-99.	
6.	Lourenco, A. P., Barry-Brooks, M., Baird, G. L., Tuttle, A., & Mainiero, M. B. (2015). Changes in recall type and patient treatment following implementation of screening digital breast tomosynthesis. Radiology, 274(2), 337-342.	
7.	McCarthy, A. M., Kontos, D., Synnestvedt, M., Tan, K. S., Heitjan, D. F., Schnall, M., et al. (2014). Screening outcomes following implementation of digital breast tomosynthesis in a general-population screening program. Journal of the National Cancer Institute, 106(11).	
8.	McDonald, E. S., McCarthy, A. M., Akhtar, A. L., Synnestvedt, M. B., Schnall, M., & Conant, E. F. (2015). Baseline Screening Mammography: Performance of Full-Field Digital Mammography Versus Digital Breast Tomosynthesis. AJR. American Journal of Roentgenology, 205(5), 1143-1148.	

#	Reference	Comments
9.	McDonald, E. S., Oustimov, A., Weinstein, S. P., Synnestvedt, M. B., Schnall, M., & Conant, E. F. (2016). Effectiveness of Digital Breast Tomosynthesis Compared With Digital Mammography: Outcomes Analysis From 3 Years of Breast Cancer Screening.[Erratum appears in JAMA Oncol. 2016 Apr;2(4):549; PMID: 26986044]. JAMA Oncology, 2(6), 737-743.	Full text unavailable
10.	Powell, J. L., Hawley, J. R., Lipari, A. M., Yildiz, V. O., Erdal, B. S., & Carkaci, S. (2017). Impact of the Addition of Digital Breast Tomosynthesis (DBT) to Standard 2D Digital Screening Mammography on the Rates of Patient Recall, Cancer Detection, and Recommendations for Short-term Follow-up. [Observational Study]. Academic Radiology, 24(3), 302-307.	
11.	Procasco, M. (2016). Comparison of Digital Breast Tomosynthesis vs Full-Field Digital Mammography in Recall Rates and Cancer Detection Rates. Radiologic Technology, 87(3), 349-351.	Full text unavailable
12.	Rafferty, E. A., Rose, S. L., Miller, D. P., Durand, M. A., Conant, E. F., Copit, D. S., et al. (2017). Effect of age on breast cancer screening using tomosynthesis in combination with digital mammography. Breast Cancer Research & Treatment, 164(3), 659-666.	
13.	Sharpe, R. E., Jr., Venkataraman, S., Phillips, J., Dialani, V., Fein-Zachary, V. J., Prakash, S., et al. (2016). Increased Cancer Detection Rate and Variations in the Recall Rate Resulting from Implementation of 3D Digital Breast Tomosynthesis into a Population-based Screening Program. Radiology, 278(3), 698-706.	
	Synthesized Digital Mammography	
1.	Ambinder, E. B., Harvey, S. C., Panigrahi, B., Li, X., & Woods, R. W. (2018). Synthesized Mammography: The New Standard of Care When Screening for Breast Cancer with Digital Breast Tomosynthesis? Academic Radiology, 25, 25	
2.	Aujero, M. P., Gavenonis, S. C., Benjamin, R., Zhang, Z., & Holt, J. S. (2017). Clinical Performance of Synthesized Two-dimensional Mammography Combined with Tomosynthesis in a Large Screening Population. Radiology, 283(1), 70-76.	
3.	Bernardi, D., Macaskill, P., Pellegrini, M., Valentini, M., Fanto, C., Ostillio, L., et al. (2016). Breast cancer screening with tomosynthesis (3D mammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study. *Lancet Oncology, 17*(8), 1105-1113	

#	Reference	Comments
4.	Caumo, F., Zorzi, M., Brunelli, S., Romanucci, G., Rella, R., Cugola, L., et al. (2017). Digital Breast Tomosynthesis with Synthesized Two-Dimensional Images versus Full-Field Digital Mammography for Population Screening: Outcomes from the Verona Screening Program. Radiology, 170745.	
5.	Hofvind, S., Hovda, T., Holen, A. S., Lee, C. I., Albertsen, J., Bjorndal, H., et al. (2018). Digital Breast Tomosynthesis and Synthetic 2D Mammography versus Digital Mammography: Evaluation in a Population-based Screening Program. Radiology, 171361	
6.	Romero Martín, S., Raya Povedano, J. L., Cara García, M., Santos Romero, A. L., Pedrosa Garriguet, M., & Álvarez Benito, M. (2018). Prospective study aiming to compare 2D mammography and tomosynthesis + synthesized mammography in terms of cancer detection and recall. From double reading of 2D mammography to single reading of tomosynthesis. [Article in Press]. European Radiology, 1-8.	
7.	Zuckerman, S. P., Conant, E. F., Keller, B. M., Maidment, A. D. A., Barufaldi, B., Weinstein, S. P., et al. (2016). Implementation of synthesized two-dimensional mammography in a population-based digital breast tomosynthesis screening program. [Article]. Radiology, 281(3), 730-736.	

Table A8. Quality assurance practices

1.	Alberdi, R. Z., Llanes, A. B., Ortega, R. A., Exposito, R. R., Collado, J. M., Verdes, T. Q., et al. (2011). Effect of radiologist experience on the risk of false-positive results in breast cancer screening programs. European Radiology, 21(10), 2083-2090.	
2.	Barlow, W. E., Chi, C., Carney, P. A., Taplin, S. H., D'Orsi, C., Cutter, G., et al. (2004). Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst, 96(24), 1840-1850.	See table 2
3.	Buist, D. S., Anderson, M. L., Haneuse, S. J., Sickles, E. A., Smith, R. A., Carney, P. A., et al. (2011). Influence of annual interpretive volume on screening mammography performance in the United States. Radiology, 259(1), 72-84.	
4.	Coldman, A. J., Major, D., Doyle, G. P., D'Yachkova, Y., Phillips, N., Onysko, J., et al. (2006). Organized breast screening programs in Canada: effect of radiologist reading volumes on outcomes. Radiology, 238(3), 809-815.	Added from other source
5.	Cornford, E., Reed, J., Murphy, A., Bennett, R., & Evans, A. (2011). Optimal screening mammography reading volumes; evidence from real life in the East Midlands region of the NHS Breast Screening Programme. Clinical Radiology, 66(2), 103-107.	
6.	Duncan, K. A., & Scott, N. W. (2011). Is film-reading performance related to the number of films read? The Scottish experience. Clin Radiol, 66(2), 99-102.	Added from other sources
7.	Théberge, I., Hébert-Croteau, N., Langlois, A., Major, D., & Brisson, J. (2005). Volume of screening mammography and performance in the Quebec population-based Breast Cancer Screening Program. CMAJ, 172(2), 195-199.	
8.	Theberge, I., Chang, S. L., Vandal, N., Daigle, J. M., Guertin, M. H., Pelletier, E., et al. (2014). Radiologist interpretive volume and breast cancer screening accuracy in a Canadian organized screening program. J Natl Cancer Inst, 106(3), djt461.	

#	Reference	Comments
	[Search Start Date: 2003]	
1.	Almazan, R., Ascunce, N., Barcos, A., Bare, M., Baroja, A., Belvis, F., et al. (2012). Effect of protocol-related variables and women's characteristics on the cumulative false-positive risk in breast cancer screening. Annals of Oncology, 23(1), 104-111.	See Roman et al. 2012
2.	Bennett, R. L., Sellars, S. J., Blanks, R. G., & Moss, S. M. (2012). An observational study to evaluate the performance of units using two radiographers to read screening mammograms. Clinical Radiology, 67(2), 114-121.	Radiographers as mammogram readers
3.	Caumo, F., Brunelli, S., Zorzi, M., Baglio, I., Ciatto, S., & Montemezzi, S. (2011). Benefits of double reading of screening mammograms: retrospective study on a consecutive series. Radiologia Medica, 116(4), 575-583.	2011a
4.	Caumo, F., Brunelli, S., Tosi, E., Teggi, S., Bovo, C., Bonavina, G., et al. (2011). On the role of arbitration of discordant double readings of screening mammography: experience from two Italian programmes. Radiol Med, 116(1), 84-91.	2011b
5.	Ciatto, S., Ambrogetti, D., Bonardi, R., Catarzi, S., Risso, G., Rosselli Del Turco, M., et al. (2005). Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening programme. Journal of Medical Screening, 12(2), 103-106.	
6.	Gromet, M. (2008). Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. AJR. American Journal of Roentgenology, 190(4), 854-859.	The article also repots on comparison between single and double reading
7.	Klompenhouwer, E. G., Voogd, A. C., den Heeten, G. J., Strobbe, L. J., de Haan, A. F., Wauters, C. A., et al. (2015). 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. European Journal of Cancer, 51(3), 391-399.	2015a
8.	Klompenhouwer, E. G., Voogd, A. C., den Heeten, G. J., Strobbe, L. J., Tjan-Heijnen, V. C., Broeders, M. J., et al. (2015). Discrepant screening mammography assessments at blinded and non-blinded double reading: impact of arbitration by a third reader on screening outcome. Multicenter Study]. European Radiology, 25(10), 2821-2829.	2015b

9.	Klompenhouwer, E. G., Weber, R. J., Voogd, A. C., den Heeten, G. J., Strobbe, L. J., Broeders, M. J., et al. (2015). Arbitration of discrepant BI-RADS 0 recalls by a third reader at screening mammography lowers recall rate but not the cancer detection rate and sensitivity at blinded and non-blinded double reading. Breast, 24(5), 601-607	2015c
10.	Liston, J. C., & Dall, B. J. (2003). Can the NHS Breast Screening Programme afford not to double read screening mammograms? Clinical Radiology, 58(6), 474-477.	
11.	Mullen, L. A., Panigrahi, B., Hollada, J., Panigrahi, B., Falomo, E. T., & Harvey, S. C. (2017). Strategies for Decreasing Screening Mammography Recall Rates While Maintaining Performance Metrics. Acad Radiol, 24(12), 1556-1560.	This article also reports on the benefits of performance feedback
12.	Posso, M. C., Puig, T., Quintana, M. J., Sola-Roca, J., & Bonfill, X. (2016). Double versus single reading of mammograms in a breast cancer screening programme: a cost-consequence analysis. European Radiology, 26(9), 3262-3271.	
13.	Salas, D., Ibanez, J., Roman, R., Cuevas, D., Sala, M., Ascunce, N., et al. (2011). Effect of start age of breast cancer screening mammography on the risk of false-positive results. Preventive Medicine, 53(1-2), 76-81.	
14.	Shaw, C. M., Flanagan, F. L., Fenlon, H. M., & McNicholas, M. M. (2009). Consensus review of discordant findings maximizes cancer detection rate in double-reader screening mammography: Irish National Breast Screening Program experience. Radiology, 250(2), 354-362.	
15.	Taylor-Phillips, S., Wallis, M. G., Jenkinson, D., Adekanmbi, V., Parsons, H., Dunn, J., et al. (2016). Effect of Using the Same vs Different Order for Second Readings of Screening Mammograms on Rates of Breast Cancer Detection: A Randomized Clinical Trial. JAMA, 315(18), 1956-1965.	
	Audit/Performance Feedback [Search Start Date: 2003]	
1.	Carney, P. A., Geller, B. M., Sickles, E. A., Miglioretti, D. L., Aiello Bowles, E. J., Abraham, L., et al. (2011). Feasibility and satisfaction with a tailored web-based audit intervention for recalibrating radiologists' thresholds for conducting additional work-up. Academic Radiology, 18(3), 369-376.	

#	Reference	Comments
2.	Carney, P. A., Abraham, L., Cook, A., Feig, S. A., Sickles, E. A., Miglioretti, D. L., et al. (2012). Impact of an educational intervention designed to reduce unnecessary recall during screening mammography. Acad Radiol, 19(9), 1114-1120.	
3.	Geertse, T. D., Holland, R., Timmers, J. M., Paap, E., Pijnappel, R. M., Broeders, M. J., et al. (2015). Value of audits in breast cancer screening quality assurance programmes. European Radiology, 25(11), 3338-3347.	
4.	Hofvind, S., Bennett, R. L., Brisson, J., Lee, W., Pelletier, E., Flugelman, A., et al. (2016). Audit feedback on reading performance of screening mammograms: An international comparison. [Comparative Study]. Journal of Medical Screening, 23(3), 150-159.	
5.	Liston, J. C., & Dall, B. J. (2003). Can the NHS Breast Screening Programme afford not to double read screening mammograms? Clinical Radiology, 58(6), 474-477.	
6.	Mullen, L. A., Panigrahi, B., Hollada, J., Panigrahi, B., Falomo, E. T., & Harvey, S. C. (2017). Strategies for Decreasing Screening Mammography Recall Rates While Maintaining Performance Metrics. Acad Radiol, 24(12), 1556-1560.	Added from other source. This article also reports on approaches to double reading.
	Comparison with Prior Mammograms [Search Start Date: 2003]	
1.	Hayward, J. H., Ray, K. M., Wisner, D. J., Kornak, J., Lin, W., Joe, B. N., et al. (2016). Improving Screening Mammography Outcomes Through Comparison With Multiple Prior Mammograms. AJR Am J Roentgenol, 1-7.	Added from other source
2.	Klompenhouwer, E. G., Duijm, L. E., Voogd, A. C., den Heeten, G. J., Strobbe, L. J., Louwman, M. W., et al. (2014). Re-attendance at biennial screening mammography following a repeated false positive recall. Breast Cancer Research & Treatment, 145(2), 429-437.	
3.	Taylor-Phillips, S., Wallis, M. G., Duncan, A., & Gale, A. G. (2012). Use of prior mammograms in the transition to digital mammography: a performance and cost analysis. European Journal of Radiology, 81(1), 60-65.	

4.	Yankaskas, B. C., May, R. C., Matuszewski, J., Bowling, J. M., Jarman, M. P., & Schroeder, B. F. (2011). Effect of observing change from comparison mammograms on performance of screening mammography in a large community-based population. Radiology, 261(3), 762-770.	Added from other source
	Number of Mammographic Views	
1.	Agbaje, O. F., Astley, S. M., Gillan, M. G. C., Boggis, C. R. M., Wilson, M., Barr, N. B., et al. (2006) Mammography reading with Computer-Aided Detection (CAD): Single view vs two views. Vol. 4046 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (pp. 125-130).	
2.	Almazan, R., Ascunce, N., Barcos, A., Bare, M., Baroja, A., Belvis, F., et al. (2012). Effect of protocol-related variables and women's characteristics on the cumulative false-positive risk in breast cancer screening. Annals of Oncology, 23(1), 104-111.	See Roman et al. 2012
3.	Salas, D., Ibanez, J., Roman, R., Cuevas, D., Sala, M., Ascunce, N., et al. (2011). Effect of start age of breast cancer screening mammography on the risk of false-positive results. Preventive Medicine, 53(1-2), 76-81.	
	Mammographic Compression [Search Start Date: 2003]	
1.	Holland, K., Sechopoulos, I., den Heeten, G., Mann, R. M., & Karssemeijer, N. (2016) Performance of breast cancer screening depends on mammographic compression. Vol. 9699. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (pp. 183-189).	
	Batch Reading of Mammograms [Search Start Date: 2003]	
1.	Burnside, E. S., Park, J. M., Fine, J. P., & Sisney, G. A. (2005). The use of batch reading to improve the performance of screening mammography. [Research Support, Non-U.S. Gov't]. AJR. American Journal of Roentgenology, 185(3), 790-796.	

#	Reference	Comments
2.	Ghate, S. V., Soo, M. S., Baker, J. A., Walsh, R., Gimenez, E. I., & Rosen, E. L. (2005). Comparison of recall and cancer detection rates for immediate versus batch interpretation of screening mammograms. [Comparative Study]. Radiology, 235(1), 31-35.	

Table A9. Radiologist characteristics

1.	Alberdi, R. Z., Llanes, A. B., Ortega, R. A., Exposito, R. R., Collado, J. M., Verdes, T. Q., et al. (2011). Effect of radiologist experience on the risk of false-positive results in breast cancer screening programs. European Radiology, 21(10), 2083-2090.	Added from other source
2.	Barlow, W. E., Chi, C., Carney, P. A., Taplin, S. H., D'Orsi, C., Cutter, G., et al. (2004). Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst, 96(24), 1840-1850.	Added from other source
3.	Carney, P. A., Elmore, J. G., Abraham, L. A., Gerrity, M. S., Hendrick, R. E., Taplin, S. H., et al. (2004). Radiologist uncertainty and the interpretation of screening. Med Decis Making, 24(3), 255-264.	Added from other source
4.	Cornford, E., Reed, J., Murphy, A., Bennett, R., & Evans, A. (2011). Optimal screening mammography reading volumes; evidence from real life in the East Midlands region of the NHS Breast Screening Programme. [Multicenter Study]. Clinical Radiology, 66(2), 103-107.	
5.	DiPrete, O., Lourenco, A. P., Baird, G. L., & Mainiero, M. B. (2018). Screening Digital Mammography Recall Rate: Does It Change with Digital Breast Tomosynthesis Experience? Radiology, 286(3), 838-844.	Added from other source
6.	Elmore, J. G., Jackson, S. L., Abraham, L., Miglioretti, D. L., Carney, P. A., Geller, B. M., et al. (2009). Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology, 253(3), 641-651.	Added from other source
7.	Miglioretti, D. L., Gard, C. C., Carney, P. A., Onega, T. L., Buist, D. S., Sickles, E. A., et al. (2009). When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology, 253(3), 632-640.	Added from other source
8.	Smith-Bindman, R., Chu, P., Miglioretti, D. L., Quale, C., Rosenberg, R. D., Cutter, G., et al. (2005). Physician predictors of mammographic accuracy. J Natl Cancer Inst, 97(5), 358-367.	Added from other source

1.	Barlow, W. E., Chi, C., Carney, P. A., Taplin, S. H., D'Orsi, C., Cutter, G., et al. (2004). Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst, 96(24), 1840-1850.	Added from other source
2.	Elmore, J. G., Jackson, S. L., Abraham, L., Miglioretti, D. L., Carney, P. A., Geller, B. M., et al. (2009). Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology, 253(3), 641-651.	
3.	Smith-Bindman, R., Chu, P., Miglioretti, D. L., Quale, C., Rosenberg, R. D., Cutter, G., et al. (2005). Physician predictors of mammographic accuracy. J Natl Cancer Inst, 97(5), 358-367.	
4.	Tan, A., Freeman, D. H., Jr., Goodwin, J. S., & Freeman, J. L. (2006). Variation in false-positive rates of mammography reading among 1067 radiologists: a population-based assessment. Breast Cancer Res Treat, 100(3), 309-318.	Added from other source
1.	Barlow, W. E., Chi, C., Carney, P. A., Taplin, S. H., D'Orsi, C., Cutter, G., et al. (2004). Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst, 96(24), 1840-1850.	Associations between litigation concerns and specificity (not recall rates). See table 4.
2.	Elmore, J. G., Taplin, S. H., Barlow, W. E., Cutter, G. R., D'Orsi, C. J., Hendrick, R. E., et al. (2005). Does litigation influence medical practice? The influence of community radiologists' medical malpractice perceptions and experience on screening mammography. Radiology, 236(1), 37-46.	

Appendix 5. Data from original studies

Table A10. Technology

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
1st Author, Date [Country]	• N/A	• Program/Study Name • Study period • Target age • Screening frequency • Sample size • Age of women	• Factor of study: Screen-film mammography vs. Full field digital mammography • Other potential influencing factors: reading approach, radiologist experience, radiologist training, and etc.	• Recall rate • False positive • Cancer detection rate • Positive predictive value	Conclusions • Author reported conclusions Limitations • Author reported limitations	• Comments (if any)
Screen-Film Mammography vs. Digital Mammography						
Campari, 2016 [Italy]	• N/A	• Program: Reggio Emilia Breast Cancer Screening Program • Study period: 1 January 2011 – 31 December 2012 • Target age: 45-74 years • Screening frequency: 45-49 years: annually 50-74 years: every 2 years	• Factor of study: Screen-film mammography vs. Digital mammography • Other potential influencing factors: **Reading Approach:** Double reading with arbitration. Second reading may be unblinded **Readers' Training:**	**Recall Rate (%)** • Screen-film: Overall: 3.3 First screen: 5.5 Subsequent screen: 2.6 • Digital: Overall: 4.4 First screen: 9.2 Subsequent screen: 3.6	**Author Reported Conclusions** • "The introduction of digital mammography in our organized screening programs led to an increased recall rate. The effect was limited to the first few months after the introduction and was attenuated by the double reading with arbitration. We	• Transition from screen-film to digital mammography occurred on January 1, 2012 • Article also reports performance measures for different age groups by mammography technology • Article also reports recall rate by

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Sample size (# of screens) = 87,436 Screen-Film: 42, 240 First: 9722 Subsequent: 32,518 Digital: 45, 196 First: 6311 Subsequent: 38, 885 • Mean age (years): 56 Screen-Film: 56.2 Digital: 55.7	Minimum of 4-6 months of digital mammography training prior to transition **Prior Mammograms:** "there was a bigger increase in the recall rate among women attending their first screening round than among those attending subsequent rounds. It is possible that the availability of previous screen film mammograms, which were consulted before making a recall decision, may have partially reduced the impact of digital mammography on the recall rate."	• Adjusted RR [Digital vs. Screen film] (95% CI) RR: 1.46 (1.37, 1.56) **Detection Rate per 1000** • Screen-film: Overall: 5.9 First screen: 4.5 Subsequent screen: 6.3 • Digital: Overall: 5.2 First screen: 5.5 Subsequent screen: 5.1 • Adjusted RR [Digital vs. Screen film] (95% CI) RR: 0.95 (0.79, 1.13) **PPV (%)** • Screen-film: Overall: 18.0 First screen: 8.3 Subsequent screen: 24.0 • Digital: Overall: 11.8 First screen: 6.0	did not observe any effect on detection rate." **Author Reported Limitations** • "our comparison does not come from a randomized design and it is therefore impossible to rule out changes in the incidence and prevalence of breast cancer during the study period." • "We did not have enough follow-up time to analyze interval cancers after digital mammography or the detection rate of advanced cancer during subsequent rounds: these are the only two indicators that can measure screening sensitivity." • "although a large number of women were included in the study, the number of cancers was small. Thus, even considerable differences in the	month and mammography technology

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				Subsequent screen: 14.3 • <u>RR [Digital vs. Screen film] (95% CI)</u> RR: 0.70 (95% CI: 0.59, 0.84) **Detection Rate for DCIS** • <u>RR [Digital vs. Screen film] (95% CI)</u> RR: 0.91 (0.59, 1.40) * adjusted RR accounts for age and screening round	sensitivity of the two screening techniques could not be detected as statistically significant and it was difficult to study the learning curve for sensitivity."	
Glynn, 2011 [USA]	• N/A	• Study: retrospective audit of performance measures before/after transition • Study period: 2004-2009 **Groups** [Screen-film] Baseline: 2004-2005 [Digital] Digital year 1: 2007 Digital year 2: 2008 Digital year 3: 2009 • Sample size (# of screens)	• Factor of study: Screen-film mammography vs. Full-field digital mammography • Other potential influencing factors: **Technology** Computer-aided detection was used **Radiologist Experience** "Each of the three radiologists is a	<u>Aggregate Results</u> **Recall Rate (%; 95% CI)** • Baseline: 6.0 (5.7, 6.3) • Digital yr 1: 7.1 (6.6, 7.6) • Digital yr 2: 8.0 (7.4, 8.7) • Digital yr 3: 8.5 (8.1, 9.0) **Cancer Detection Rate (per 1000 women; 95% CI)**	**Author Reported Conclusions** • "In conclusion, our experience has been that moving from analog technology to digital screening technology increased recall and cancer detection rates in the first few years after the transition. PPV 1 and PPV 3 were particularly reduced with respect to calcifications, a	• Transition from analog to full-field digital mammography occurred in in November 2006

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		Technology Screen-film: 32,600 Digital: 33,879 **Study Period** Baseline: 32,600 Digital year 1: 11,358 Digital year 2: 7,924 Digital year 3: 14,597 • Sample size (# of radiologists): 3 • Lesion analysis sample size (# of lesions) Baseline: 1,859 Digital year 1: 959 Digital year 2: 675 • Median age (range): 52 years (27-92 years)	Mammography Quality Standards Act–certified dedicated breast imager with at least 10 years of experience in breast imaging." **Radiologist Training** "The radiologists had completed training in digital mammography in compliance with the Mammography Quality Standards Act."	• Baseline: 3.34 (2.75, 4.03) • Digital yr 1: 5.28 (4.03, 6.80) • Digital yr 2: 5.93 (4.36, 7.89) • Digital yr 3: 4.52 (3.50, 5.75) **PPV$_1$ ["probability of cancer after positive mammographic interpretation"] (%, 95% CI)** • Baseline: 5.6 (4.6, 6.7) • Digital yr 1: 7.5 (5.7, 9.6) • Digital yr 2: 7.4 (5.4, 9.8) • Digital yr 3: 5.3 (4.1, 6.7) **PPV$_3$ ["probability of cancer among patients undergoing biopsy after…(BI-RADS) assessment of 4 or 5"] (%; 95% CI)** • Baseline: 44.5 (36.5, 53.7) • Digital yr 1: 31.3 (23.9, 40.2) • Digital yr 2: 38.2 (28.1, 50.8)	finding that is supported by the findings of other recent studies" **Author Reported Limitations** • "we assumed that the overall screening population remained stable for the years being studied. Without prior data for comparison, the recall rate in a group with a relatively greater number of new screening studies may be higher. In our study, this could be an additional factor leading to a perceived increased recall rate due to new technology, when in fact the increased number of new screens played a role." • "PPVs are highly dependent on the proportion of subjects who have the disease (prior probability of disease) and may be	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				• Digital yr 3: 30.3 (23.4, 38.5) Calcified Lesions **Recall Rate (%; 95% CI)** • Baseline: 13.8 (12.2, 15.6) • Digital yr 1: 23.9 (20.9, 27.2) • Digital yr 2: 17.9 (14.9, 21.4) **PPV$_1$ (%; 95% CI)** • Baseline: 15.2 (10.8, 20.7) • Digital yr 1: 10.5 (6.7, 15.6) • Digital yr 2: 11.6 (6.3, 19.4) **PPV$_3$ (%, 95% CI)** • Baseline: 41.1 (29.2, 56.1) • Digital yr 1: 21.8 (14.0, 32.5) • Digital yr 2: 24.1 (13.2, 40.5) Noncalcified Lesions **Recall Rate (%; 95% CI)** • Baseline: 86.2 (82.0, 90.5) • Digital yr 1: 76.1 (70.7, 81.9)	different in different clinical settings.” • “we did not collect breast glandular density data in this study.” • “this was a single-institution study rather than a multicenter study, and the findings reflect the experience of only three high-volume radiologists.”	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				- Digital yr 2: 82.1 (75.4, 89.2) **PPV$_1$ (%; 95% CI)** - Baseline: 4.9 (3.9, 6.1) - Digital yr 1: 5.9 (4.3, 7.9) - Digital yr 2: 4.3 (2.8, 6.4) **PPV$_3$ (%; 95% CI)** - Baseline: 52.3 (41.4, 65.2) - Digital yr 1: 43.4 (31.4, 58.5) - Digital yr 2: 44.4 (28.5, 66.1)		
Karssemeijer, 2009 [Netherlands]	- N/A	- Program: population-based breast cancer screening program at the Preventicon screening centre - Study period: Up to 5 years following program start date (September 2003) - Target age: 50-75 years - Screening frequency: 2-year interval	- Factor of study: Full-field digital mammography (FFDM) with computer-aided diagnosis (CAD) vs. Screen-film mammography (SFM) - Other potential influencing factors: **Mammographic Views** Initial examinations:	**Recall Rate (%)** - Initial screen SFM: 2.32 FFDM: 4.41 p: <0.001 - Subsequent screen SFM: 1.17 FFDM: 1.70 p: <0.001 **Cancer Detection Rate (%)** - Initial screen SFM: 0.62 FFDM: 0.77	**Author Reported Conclusions** - "Results indicate that with the FFDM-CAD combination and double reading, the detection is as good as that with SFM, and detection of clustered microcalcifications and DCIS is improved with FFDM using CAD." **Author Reported Limitations**	- Article also reports recall rates and PPV per lesion type (mass, architectural distortion, clustered microcalcifications, and other)

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Sample size (# of screens) = 367,600 **Technology** FFDM: 56,518 SFM: 311,082 **Initial Procedures** FFDM: 10,307 SFM: 38,754	Two views acquired (craniocaudal and mediolateral oblique). <u>Subsequent examinations:</u> Mediolateral oblique views acquired. Craniocaudal views acquired when indicated by breast density and visible abnormality criteria. **Radiographers' Training** Extensive training in FFDM use received. "They were instructed to obtain the best possible positioning and compression with each modality..." **Technology** "a dedicated workstation with a high-resolution monitor was installed…to allow	p: 0.11 • <u>Subsequent screen</u> SFM: 0.49 FFDM: 0.55 p: 0.12 **Invasive Cancers Rate (%)** • <u>Initial screen</u> SFM: 0.49 FFDM: 0.54 p: 0.46 • <u>Subsequent screen</u> SFM: 0.40 FFDM: 0.40 p: 0.96 **Ductal Carcinoma in Situ (DCIS) Rate (%)** • <u>Initial screen</u> SFM: 0.12 FFDM: 0.22 p: 0.015 • <u>Subsequent screen</u> SFM: 0.08 FFDM: 0.12 p: 0.007 **PPV of Recall (%)** • <u>Initial screen</u> SFM: 26.8 FFDM: 17.4 • <u>Subsequent screen</u>	• "the contribution of FFDM and CAD could not be evaluated separately because they were introduced at the same time." • "the unavailability of detailed pathology reports, which prohibited reliable analysis of the histologic grades of DCIS." • "The study was not designed as a randomized controlled trial. Assignment of modality was determined according to availability, which was random, and also according to the previous screening, as women who once had undergone FFDM remained in the digital track. As in the initial phase, all FFDM screenings were initial screenings, this led to a slight bias toward younger women being assigned to FFDM. This was visible as a small bias in mean	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			proper viewing of digital mammograms." Soft-copy reading used with FFDM. CAD was used with FFDM but not SFM. **Reading Approach** Independent double reading with consensus. "Mammograms were interpreted in a batch mode within 2 days of acquisition." **Radiologist Experience** "two radiologists… [had] more than 15 years of experience in mammography screening". "All radiologists… had more than 2 years experience with working in a digital radiology environment... None of	SFM: 43.1 FFDM: 30.4	age in the two groups...Bias would be in favor of SFM, since incidence increases with age" • "we mention the effect of multiple screening rounds on the expected screening outcome. When more early-stage cancers are found with FFDM using CAD, this will lead to less-invasive cancers in subsequent FFDM screenings and less interval cancers. Because of incomplete data on interval cancers, we could not investigate this issue."	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			the readers had experience with use of FFDM in screening or with the type of processing implemented in the FFDM system used in the study. All radiologists had extensive experience with clinical use of digital mammography with a computed radiography detector.” **Reading Volume** Other than the two radiologists with >15 years of experience, seven radiologists who had reading volumes of >5,000 screens/year conducted the remainder of readings **Prior Mammograms**			

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Prior mammograms were used in subsequent screenings			
Vernacchia, 2009 [USA]	• N/A	• Study: Small community-based radiology practice • Study period: **Audit Periods** [Screen-film] <u>Audit 1</u>: July 1, 2004-June 30, 2005 [Digital] <u>Audit 2</u>: September 1, 2005-August 31, 2006 <u>Audit 3</u>: September 1, 2006-August 31, 2007 <u>Audit 4</u>: September 1, 2007-August 31, 2008 • Sample size (# of screens) **Audit Period** [Screen-film] Audit 1: 4,838 [Digital] Audit 2: 6,875 Audit 3: 7,379 Audit 4: 7,294	• Factor of study: Screen film mammography vs. Digital mammography • Other potential influencing factors: **Radiologist Experience** Two radiologist part of audits 1-3 had >20 years experience in screening mammography **Radiologist Training** A radiologist part of audits 2-4 "recently graduated and board-certified with only residency training in mammography". "Except for training in digital	**Recall Rate for BI-RADS category 0 (%)** • Audit 1: 5.9 • Audit 2: 10.2 • Audit 3: 7.5 • Audit 4: 9.0 **Cancer Detection Rate (per 1,000 women screened)** • Audit 1: 4.1 • Audit 2: 7.9 • Audit 3: 5.1 • Audit 4: 6.9	**Author Reported Conclusion** • "In this community-based mammography practice, an increase in the cancer detection rate occurred initially during the conversion from screen-film to digital mammography, which subsequently decreased but remained higher than before digital conversion. This study suggests that the new technology alone is responsible for the increased number of cancers detected in patients with dense breasts that were not previously found using screen-filming"	• Article also report recall and cancer detection rates between radiologists 1 and 2 (permanent staff)

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			mammography as required under the Mammography Quality Standards Act, none of these physicians had had any prior experience with the interpretation of digital mammography." **Guidelines** "all imaging studies fol- lowed ACR (American College of Radiology) and ACS (American Cancer Society) guidelines for screening mammography."		**Author Reported Limitations** - "One shortcoming of our study, because of the limitations of our mammography reporting and tracking software, was our inability to extract information on the overall breast density of our patient population. As a result, we are unable to determine whether our increased cancer detection rate is due to a higher than normal population of patients with dense breasts." - "Also, our database does not allow us to detect how many cancers were detected in patients we saw for the first time during audit 2."	
Sala, 2015 [Spain]	- N/A	- Program: Retrospective cohort study of women in population-based breast cancer screening program in Barcelona	- Factor of study: Screen-film mammography (SFM) vs. Full-field digital mammography (FFDM)	**Recall Rate (%)** - Overall Screen-film: 5.57 Digital: 4.20 p: <0.001 - Initial screen	**Author Reported Conclusions** - "Digitalization has supposed an improvement in early diagnosis because	- Screening occurred in two radiology units, where transition from SFM to FFDM occurred in 2007 and 2004.

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Study period: **Research Units (RU)** RU1 – SFM: January 1998 – March 2007 RU1 – FFDM: March 2007 – December 2010 RU2 – SFM: January 2001 – September 2004 RU2 – FFDM: September 2004 – December 2010 • Target age: 50-69 years • Screening frequency: 2-year interval • Sample size (# of screens) **Technology** SFM: 82,961 FFDM: 79,031	• Other potential influencing factors: **Mammographic Views** Two views for each breast (Mediolateral oblique and craniocaudal views) **Reading Approach** Double reading with arbitration. **Prior Mammograms** Always available for successive screenings **Guidelines** "The program was based on the European Guidelines for Quality Assurance in Mammographic Screening [14] and its results met the Europe Against Cancer standards."	Screen-film: 11.00 Digital: 11.73 p: 0.032 • Successive screen Screen-film: 3.72 Digital: 2.50 p: <0.001 **Cancer Detection Rate (%)** • Overall Screen-film: 0.42 Digital: 0.43 p: 0.685 • Initial screen Screen-film: 0.39 Digital: 0.55 p: 0.024 • Successive screen Screen-film: 0.43 Digital: 0.40 p: 0.503 • Adjusted OR (95% CI) *1st SFM period:* reference *2nd SFM period:* 0.84 (0.61, 1.16) *3rd SFM period:* 1.20 (0.89, 1.61) *4th SFM period:* 1.01 (0.74, 1.37) *1st FFDM period:* 1.06 (0.77, 1.44) *2nd FFDM period:*	DCIS and small invasive cancers increased without a change in detection rate. Moreover, false-positive reduction without an increase in the interval cancer rate was confirmed." **Author Reported Limitations** • "Although the study period is one of the longest ever analysed, the number of DCIS does not allow exploration of trends in tumoral grade during the digital period." • "We had no information on interval cancer subtypes, and therefore we could not assess the behaviour of false negatives before and after the shift to digital technology."	• Article also provides performance measures by study period

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				1.19 (0.88, 1.62) *3rd FFDM period:* 1.04 (0.75, 1.42) *4th FFDM period:* 1.14 (0.84, 1.55) **Invasive Carcinomas Detection Rate (%)** • Overall Screen-film: 0.36 Digital: 0.33 p: 0.462 • Initial screen: Screen-film: 0.31 Digital: 0.42 p: 0.091 • Successive screen: Screen-film: 0.37 Digital: 0.31 p: 0.089 • Adjusted OR (95% CI) for invasive cancers *1st SFM period:* Reference *2nd SFM period:* 0.80 (0.57, 1.12) *3rd SFM period:* 1.12 (0.82, 1.53) *4th SFM period:* 0.97 (0.70, 1.34) *1st FFDM period:* 1.00 (0.71, 1.40) *2nd FFDM period:* 1.05 (0.75, 1.46)		

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				3rd FFDM period: 0.86 (0.61, 1.22) *4th FFDM period:* 0.97 (0.69, 1.36) **In Situ Carcinoma Detection Rate (%)** • Overall Screen-film: 0.05 Digital: 0.09 p: 0.010 • Initial screen: Screen-film: 0.06 Digital: 0.12 p: 0.031 • Successive screen Screen-film: 0.05 Digital: 0.08 p: 0.063 • Adjusted OR (95 % CI) for DCIS *1st SFM period:* Reference *2nd SFM period:* 0.93 (0.36, 2.43) *3rd SFM period:* 2.05 (0.91, 4.63) *4th SFM period:* 1.34 (0.55, 3.26) *1st FFDM period:* 1.58 (0.65, 3.80) *2nd FFDM period:* 2.53 (1.13, 5.69) *3rd FFDM period:*		

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				2.51 (1.12, 5.66) *4th FFDM period:* 2.68 (1.20, 6.01) **Positive Predictive Value (%)** • Overall Screen-film: 8.00 Digital: 11.25 p: <0.001 • Initial screen Screen-film: 4.20 Digital: 6.43 p: 0.010 • Successive screen Screen-film: 11.14 Digital: 14.64 p: 0.004 * Odds ratios adjusted for radiology unit, age, screening round of diagnosis (initial/successive)		
Hambly, 2009 [Ireland]	• N/A	• Program: Irish National Breast Screening Program (INBSP) • Study period: January 1, 2005 – December 31, 2007 • Target age: 50-64 years	• Factor of study: Full-field digital mammography (FFDM) vs. Screen-film mammography (SFM) • Other potential influencing factors:	**Recall Rate (%)** • All screens SFM: 3.1 FFDM: 4.0 p: <0.001 • First screen SFM: 5.7 FFDM: 7.3 p: <0.001	**Author Reported Conclusions** • "FFDM resulted in significantly higher cancer detection and recall rates than screen-film mammography in women 50–64 years	• Article also reports information, including recall rate, cancer detection rate, and PPV_1, from seven other studies • Women were assigned SFM or FFDM based on the

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Screening frequency: every 2 years • Sample size (# of screens) = 188,823 **Technology** FFDM: 35,204 SFM: 153,619 **Screening Round** Initial: 53,702 Subsequent: 135,121 • Average age (years): **First Screening** FFDM: 53.5 SFM: 54.1 **Subsequent Screening** FFDM: 58.6 SFM: 58.5	**Reading Approach** Unblinded double reading with consensus "The consensus meeting was held twice weekly. All radiologists were invited to attend, and a minimum of two was required. All cases with a discrepancy in R category from the previous week were reviewed, and a consensus was reached as to whether the patient should be recalled for assessment or listed for routine screening." **Radiologist Experience** At least 5 years of experience in mammography reading **Reading Volume**	• Subsequent screen SFM: 2.0 FFDM: 2.8 p: <0.001 **Cancer Detection Rate (per 1,000 screenings)** • All screens SFM: 5.2 FFDM: 6.3 p: 0.01 • First screen SFM: 7.0 FFDM: 7.9 p: 0.483 • Subsequent screen SFM: 4.4 FFDM: 5.7 p: 0.008 **Invasive Cancer Detection Rate (per 1,000 screenings)** • All screens SFM: 4.2 FFDM: 5.0 p: 0.054 • First screen SFM: 5.7 FFDM: 6.4 p: 0.63 • Subsequent screen SFM: 3.6	old. The PPVs of FFDM and screen-film mammography were comparable. The results of this study suggest that FFDM can be safely implemented in breast cancer screening programs." **Author Reported Limitations** • "because the women screened in 2006 and 2007 have not yet had their 2-year follow-up, early false-negative studies cannot be excluded." • "Some women (~ 25%) underwent two screening mammography examinations during the study period. We do not think that this would have influenced the results of the study because women are removed from the screening population to a symptomatic service once they are diagnosed with cancer and were recalled for a specific	check-in time: "every third or fourth patient was assigned to digital mammography depending on the screening center." • Article also reports recall and cancer detection rates for mammography technology by age groups • Article also reports cancer detection rates for mammography technology by type of abnormality detected

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			20,000 examinations/year (average) **Prior Mammograms** Previous mammograms were used	FFDM: 4.4 p: 0.047 **Ductal Carcinoma In Situ [DCIS] (per 1,000 screenings)** • All screens SFM: 0.95 FFDM: 1.3 p: 0.072 • First screen SFM: 1.3 FFDM: 1.5 p: 0.66 • Subsequent screen SFM: 0.8 FFDM: 1.2 p: 0.036 **PPV$_1$ ["number of cancers detected as a percentage of the women recalled to assessment"] (%)** • All screens SFM: 16.7 FFDM: 15.7 p: 0.383 • First screen SFM: 12.2 FFDM: 10.8 p: 0.337 • Subsequent screen	abnormality only once." • "Assignment to digital or analog mammography was not influenced by the type of mammography examination previously performed and was based only on the time of check-in." • "the possibility of bias introduction during randomization" • "Information regarding breast density, menopausal status, and other risk factors such as hormone replacement therapy use, parity, and age of menarche is not recorded by the INBSP. This is a limitation of our study, and subtle differences between the two groups cannot be definitively excluded."	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				SFM: 21.9 FFDM: 20.5 p: 0.455		
Sankatsing, 2018 [Netherlands]	• N/A	• Program: Dutch breast cancer screening programme (BCSP) • Study period: 2004 – 2011 • Target age: 50 – 74 • Screening frequency: biennially • Sample size (# of screens) = 7,343,327 • Age range (years): 49-74	• Factor of study: Screen-film mammography (SFM) vs. Digital mammography (DM) • Other potential influencing factors: **Reading Approach** Double reading with consensus or arbitration **Mammographic Views** Initial screen: 2 views Subsequent screen: proportion of screens with 2 views was about 50% in 2004 and 93% in 2010	**Recall Rate (per 1,000 screens; 95% CI)** • DM: 21.0 (20.8, 21.2) • SFM: 16.0 (15.9, 16.1) **Detection Rate [all] (per 1,000 screens; 95% CI)** • DM: 6.2 (6.1, 6.3) • SFM: 5.4 (5.3, 5.4) **Detection Rate Ductal Carcinomas In Situ [DCIS] (per 1,000 screens; 95% CI)** • DM: 1.1 (1.1, 1.2) • SFM: 0.83 (0.81, 0.86) **Detection Rate Invasive (per 1,000 screens; 95% CI)** • DM: 5.1 (5.0, 5.2) • SFM: 4.5 (4.5, 4.6) **Positive Predictive Value (%; 95% CI)** • DM: 31.5 (31.1, 31.9) • SFM: 34.9 (34.5, 35.2)	**Author Reported Conclusions** • "During the transition from SFM to DM, there was a significant rise in DR [detection rate] and a stable ICR [interval cancer rates], leading to increased programme sensitivity. Although the recall rate increased, programme specificity remained high compared to other countries. These findings indicate that the performance of DM in a nationwide screening programme is not inferior to, and may be even better, than that of SFM." **Author Reported Limitations** • "Single screening examinations were not labelled as DM or	• "Screening examinations…were subdivided in initial screens, regular subsequent screens within 2.5 years after previous screening and irregular subsequent screens 2.5 years or later after previous screening...The latter were not used in this study…" • Article also provides figures of trends over time

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				**Note: above results were adjusted for age*	SFM at time of screening and information about the proportion DM and SFM, during the years in which both modalities were used, had to be obtained from the screening units. The screens for which it was uncertain whether they were performed using screen-film or digital mammography were added to the screen-film group. This could lead to underestimation of detection rates for DM and to increased apparent detections rates for SFM. The difference in detection of DM relative to SFM could thus be (somewhat) greater than we report and our estimates may therefore be conservative." • "In addition, 2% of all breast cancers in the NCR [Netherlands Cancer Registry] database could not	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					be classified as screen-detected or interval cancer."	
de Munck, 2016 [Netherlands]	- N/A	- Program: Dutch breast cancer screening program (region North-Netherlands) - Study period: 2004 – 2010 - Target age: 50 – 75 years - Screening frequency: biennial - Sample size (# of screens) = 902,868	- Factor of study: Screen-film mammography (SFM) vs. Full-field digital mammography (FFDM) - Other potential influencing factors: **Training** "All mammographic examinations are performed by specialised radiographers." **Mammographic Views** Initial screen: Two views obtained (craniocaudal and mediolateral oblique) Subsequent screen: Mediolateral oblique views obtained. Criteria used to indicate	**Recalled Women (%)** - Overall SFM: 1.26 FFDM: 1.34 p: 0.002 - Initial screen SFM: 2.07 FFDM: 3.02 p: <0.001 - Subsequent screen SFM: 1.15 FFDM: 1.14 p: 0.532 **Screen Detected Breast Cancers (per 1,000 women screened)** - Overall SFM: 5.28 FFDM: 5.24 p: 0.800 - Initial screen SFM: 5.28 FFDM: 6.01 p: 0.159 - Subsequent screen SFM: 5.28 FFDM: 5.14	**Author Reported Conclusions** - "FFDM resulted in similar rates of screen-detected and interval cancers, indicating that FFDM performs as well as SFM in a breast cancer screening program. No signs of an increase in low-grade DCIS (which might connote possible overdiagnosis) were seen. Nonetheless, after initial screening, which accounts for 12% of all screens, FFDM resulted in higher recall rate and lower PPV that requires attention." **Author Reported Limitations** - "A limitation in comparing results with other studies might lie in the fact that the Dutch screening program	- Article also provides figures with recall and detection rates by mammography technology over time

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			when craniocaudal views are obtained. **Reading Approach** Mammograms read in batch mode. Independent double reading with consensus. **Prior Mammograms** Prior mammograms available during subsequent screens	p: 0.445 **Screen Detected DCIS (per 1,000 women screened)** - Overall SFM: 0.76 FFDM: 0.85 p: 0.137 - Initial screen SFM: 0.86 FFDM: 1.18 p: 0.137 - Subsequent screen SFM: 0.74 FFDM: 0.81 p: 0.298 **Screen Detected Invasive Cancers (per 1,000 women screened)** - Overall SFM: 4.53 FFDM: 4.39 p: 0.369 - Initial screen SFM: 4.42 FFDM: 4.83 p: 0.385 - Subsequent screen SFM: 4.54 FFDM: 4.33	invites women 50–75 years of age, which differs from other screening programs mostly offering screening to women 50–69 years old. This might limit the generalisability of our results.” - “Second, in The Netherlands, screening examinations are independently read by two radiologists and – particularly in the Northern region – recall rates are relatively low. Comparing FFDM with SFM might therefore lead to different conclusions in comparison with other studies. Furthermore, a Dutch study found variation in recall rate, with less variation in detection rate, between regions in The Netherlands ...” - “Finally, during this study period a policy change towards making standard	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				p: 0.208 • **Positive Predictive Value for Screen Detected Breast Cancers (%)** • Overall SFM: 41.8 FFDM: 39.0 p: 0.004 • Initial screen SFM: 25.6 FFDM: 19.9 p: 0.002 • Subsequent screen SFM: 45.7 FFDM: 45.2 p: 0.638	craniocaudal views at subsequent screening examinations started in The Netherlands. Also, during this period some radiologists synchronous read mammograms made using FFDM and SFM. Both effects could have influenced recall or detection rate for both SFM and FFDM. However, in our data we did not find an increased recall or detection rate for SFM after subsequent screening examination, indicating that these effects were negligible."	
Theberge, 2016 [Canada]	• N/A	• Program: Quebec Breast Cancer Screening Program (Programme Québécois de Dépistage du Canada du Sein [PQDCS] • Study period: January 1, 2007 to September 30, 2012 • Target age: 50 – 69 years	• Factor of study: Digital mammography vs. Screen-film mammography (SFM) • Other potential influencing factors: **Mammographic Views** Two views (craniocaudal and	**Recall Rate (%)** • SFM: 9.0 • CR: 9.6 • DR: 13.4 **Adjusted Odds Ratio for Recall Rate (95% CI)[a]** • SFM: 1.00 • CR: 1.03 (1.01, 1.06) • CR-Fuji: 1.05 (1.02, 1.07)	**Author Reported Conclusions** • "In conclusion, this study suggests that, in the PQDCS, CR is associated with similar detection rate and a small increase in recall rate compared to SFM. This study also suggests that screening programs	• "Two types of digital technology…: computed radiography (CR) or digital direct radiography (DR). • "we cannot determine which unit produced each screening mammogram if there is more than 1 unit. Therefore,

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Screening frequency: biennially • Sample size (# of screens) = 1,585,272 **Technology** SFM: 782,894 CR: 672,125 DR: 60,023 Mixed: 70,230	mediolateral oblique views) **Reading Approach** Single reading **Reading Volume** At least 500 mammograms/year during study period **Radiologist Gender** Male (n; %): 227 (63.9) **Quality-Control Program** "quality-control program…includes regular tests of technical quality to ensure that the mammography unit, processor, and all related equipment are working properly" **Certification and Accreditation** "Centres must also be certified by the Laboratoire de Santé Publique du Québec (LSPQ)…This	• CR-Kodak: 1.02 (0.97, 1.08) • CR-Agfa: 0.93 (0.89, 0.98) • DR: 1.25 (1.19, 1.30) **Crude Detection Rates (cancers per 1,000 screens)** • SFM: 5.1 • CR: 5.1 • DR: 5.9 **Adjusted Odds Ratio for Detection Rate (95% CI)[a]** • SFM: 1.00 • CR: 0.95 (0.88, 1.03) • CR-Fuji: 0.97 (0.89, 1.05) • CR-Kodak: 0.88 (0.74, 1.05) • CR-Agfa: 0.91 (0.77, 1.08) • DR: 1.06 (0.89, 1.25) **Adjusted Odds Ratio for Invasive Detection Rate (95% CI)[a]** • SFM: 1.00 • CR: 0.95 (0.87, 1.03) • CR-Fuji: 0.95 (0.87, 1.04) • CR-Kodak: 0.92 (0.77, 1.11)	offering CR should monitor the performance taking the CR plate reader manufacturer into consideration. In our screening program, implementation of DR is associated with a detection rate similar to that of SFM, but with a significantly higher recall rate. If this situation persists, the adoption of DR may increase the adverse effects of screening with little or no benefit for women." **Author Reported Limitations** • "Our results concerning DR should be taken with caution because the conversion to DR is relatively recent in our program. We do not know if the findings observed will be the same when more centres implement DR or when follow-up is longer."	the mammograms performed in the centres using 2 different technologies over the same period…form the category we called ‘‘mixed.””

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			certification is based on annual examination by a physicist of the installations, the equipment as well as technical image quality…" "…centres must also be accredited by the Mammography Accreditation Program of the Canadian Association of Radiologists where both technical aspects and clinical image quality are evaluated. " "Certification by LSPQ and accreditation by the Canadian Association of Radiologists of a centre must be obtained for each of its mammography units as some	• CR-Agfa: 0.90 (0.75, 1.08) • DR: 1.06 (0.88, 1.27) **Adjusted Odds Ratio for Ductal Carcinoma In Situ [DCIS] detection Rate (95% CI)[a]** • SFM: 1.00 • CR: 0.93 (0.79, 1.10) • CR-Fiji: 0.97 (0.82, 1.16) • CR-Kodak: 0.70 (0.48, 1.02) • CR-Agfa: 0.91 (0.64, 1.30) • DR: 1.06 (0.74, 1.52) **Adjusted Odds Ratio for Positive Predictive Value (95% CI)[a]** • SFM: 1.00 • CR: 0.93 (0.85, 1.01) • CR-Fuji: 0.94 (0.86, 1.03) • CR-Kodak: 0.81 (0.67, 0.98) • CR-Agfa: 1.00 (0.84, 1.19) • DR: 0.88 (0.74, 1.05) [a] "Adjusted for characteristics of	• "we could not study the association between DR and performance indicators according to the manufacturer of DR units due to low number of DR mammograms." • "some misclassification of mammograms as SFM, CR, or DR may have occurred. To minimize this misclassification, we have grouped mammograms done in centres using 2 different technologies in the same period. This group of mammograms (mixed) are included in the statistical model, but the results are not shown for that group."	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			centres have more than 1 unit."	women (age, breast density, body mass index, first-degree relatives family history of breast cancer, menopausal status, parity, use of hormone replacement therapy, previous breast aspiration or biopsy, initial examination or rescreening, and year of the Quebec Breast Cancer Screening Program mammogram), radiologists (gender, year of graduation, medical school attended, and annual program screening volume), and facilities (recall rate in 2006, detection rate in 2006, facility type, and annual programme screening volume)."		
Vinnicombe, 2009	- N/A	- Program: Central and East London Breast Screening Service (CELBSS)	- Factor of study: Full-field digital mammography (FFDM) vs. Screen-	**Recall Rate [per 100 screening**	**Author Reported Conclusions**	- Article also reported crude relative risks for cancer detection

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
[United Kingdom]		• Study period: January 1, 2005 – June 30, 2007 • Target age: ≥ 50 years • Screening frequency: every 3 years • Sample size (# of screens) = 40,198 **Technology** FFDM: 8478 SFM: 31,720 • Median age in years (interquartile range): 58.0 (53.5, 63.5)	film mammography (SFM) • Other potential influencing factors: **Reading Approach** Unblinded double reading with arbitration. FFDM used hard-copy image reading. **Mammographic Views** Two views **Quality Assurance** "All mammography units …and are subjected to rigorous quality control procedures as specified in the NHSBSP (4) and European quality assurance guidelines (5)." **Radiologist Experience** >10 years of experience in beast screening (except 1 of 6 radiologists)	**mammograms] (95% CI)** • ≤ 60 years SFM: 5.03 (4.72, 5.34) FFDM: 5.23 (4.62, 5.83) • > 60 years SFM: 3.54 (3.22, 3.86) FFDM: 4.08 (3.39, 4.76) • All ages SFM: 4.43 (4.20, 4.65) FFDM: 4.79 (4.33, 5.24) **Adjusted Relative Risk for Recall rate [FFDM vs. SFM] (95% CI)** • ≤ 60 years 0.93 (0.79, 1.06) p: 0.31 • > 60 years 1.01 (0.80, 1.23) p: 0.91 • All ages 0.95 (0.84, 1.07) p: 0.44 **Cancer Detection Rate [per 100 screening**	• "Within a routine screening program, FFDM with hard-copy image reading performed as well as SFM in terms of process indicators; the meta-analysis was consistent with FFDM yielding detection rates at least as high as those for SFM." **Author Reported Limitations** • "In our study all screening mammograms, whether from SFM or FFDM, were viewed with identical conditions. For the first 18 months of the study, it was not possible to print hard copy of the optimally postprocessed GE Healthcare images ("Premium View"), and FFDM screening images obtained during this period resembled analogue film images. Thus, some cancers may have been missed at FFDM in this group.	rates, recall rates, and PPV of breast cancers among recalls • Article also performed a systematic review

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Reading Volume At least 5,000 screening mammograms/year	**mammograms] (95% CI)** • ≤ 60 years SFM: 0.52 (0.42, 0.62) FFDM: 0.61 (0.40, 0.82) • > 60 years SFM: 0.84 (0.68, 1.00) FFDM: 0.81 (0.50, 1.12) • All ages SFM: 0.65 (0.56, 0.73) FFDM: 0.68 (0.51, 0.86) **Adjusted Relative Risk for Cancer Detection Rate [FFDM vs. SFM] (95% CI)** • ≤ 60 years 1.05 (0.59, 1.51) p: 0.83 • > 60 years 0.86 (0.48, 1.25) p: 0.52 • All ages 0.95 (0.65, 1.25) p: 0.74	This did not apply to the hard-copy images from screening mammograms obtained with the Selenia unit (Lorad), which were indistinguishable from the soft-copy images." • "Cohort studies, such as ours, are more prone to be affected by confounding than paired studies or large randomized trials. In cohort studies, distinct groups of women underwent FFDM and SFM, with the allocation to screening modality being determined nonrandomly by the women themselves or by the screening program. Thus, women who underwent FFDM might have differed from those who underwent SFM in relation to factors that influence detection rates."	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				PPV of Breast Cancers Among Recalls [%] (95% CI) • ≤ 60 years SFM: 10.29 (8.36, 12.22) FFDM: 11.64 (7.85, 15.43) • > 60 years SFM: 23.67 (19.75, 27.59) FFDM: 19.85 (13.02, 26.68) • All ages SFM: 14.60 (12.75, 16.45) FFDM: 14.29 (10.88, 17.69) **Adjusted Relative Risk for PPV of Breast Cancers Among Recalls [FFDM vs. SFM] (95% CI)** • ≤ 60 years 1.07 (0.65, 1.49) p: 0.72 • > 60 years 0.84 (0.51, 1.18) p: 0.39 • All ages 0.95 (0.68, 1.23) p: 0.75	• "Although attempts were made to minimize confounding at the design and analysis stages, one cannot exclude the possibility that the findings from cohort studies might have been affected by residual or unknown confounding."	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				*Adjusted covariates for relative risk results above include age, ethnicity, referral type, and area of residence		
Hofvind, 2014 [Norway]	• N/A	• Program: Norwegian Breast Cancer Screening Program (NBCSP) • Study period: 1996 – 2010 • Target age: 50-69 years • Screening frequency: every 2 years • Sample size (# of screens) =1,837,360 **Technology** SFM: 1,391,188 FFDM: 446,172 • Median age in years (range): 58 (50 – 69)	• Factor of study: Screen-film mammography (SFM) vs. Full-field digital mammography (FFDM) • Other potential influencing factors: **Reading Approach** Independent double reading with arbitration or consensus. "If one or both radiologists have given a score of 2 or higher, a consensus or arbitration meeting … is used to determine whether to call the woman back for further assessment (recall) or not."	**Recall for further assessment (per 1,000 examinations)** • Overall SFM: 0.34 FFDM: 0.29 p: <0.001 • Baseline examinations SFM: 0.50 FFDM: 0.62 • Subsequent examinations SFM after SFM: 0.26 FFDM after SFM: 0.23 FFDM after FFDM: 0.21 **Screening-Detected Cancer Total (per 1,000 examinations)** • Overall SFM: 0.56 FFDM: 0.52 p: 0.005	**Author Reported Conclusions** • "After the initial transitional phase from SFM to FFDM, population-based screening with FFDM is associated with less harm because of lower recall and biopsy rates and higher positive predictive values after biopsy than screening with SFM." **Author Reported Limitations** • "Because the baseline characteristics of Norwegian women may be more homogeneous than those observed in other countries, our findings may not be generalizable to	• Article also provided figures with recall rate, screening-detected breast cancer rate, and interval breast cancer rate by time after implementation of FFDM • Article also reported baseline and subsequent examinations performance measures for technology by age groups

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Guidelines Follows European screening guidelines **Mammographic Views** Two views **Radiologist Training** "All radiologists working in the NBCSP are experienced in diagnostic mammography, but no systematic training or education was required before they started to read FFDM." **Reading Volume** Average of 3,600 screening mammography examinations/year between 1996 – 2005	• <u>Baseline examinations</u> SFM: 6.69 FFDM: 6.35 • <u>Subsequent examinations</u> SFM after SFM: 5.03 FFDM after SFM: 5.01 FFDM after FFDM: 4.93 **Screen-Detected Cancer for Ductual Carcinoma In Situ [DCIS] (per 1,000 examinations)** • <u>Overall</u> SFM: 0.09 FFDM: 0.11 p: 0.019 • <u>Baseline examinations</u> SFM: 1.16 FFDM: 1.50 p: <0.05 • <u>Subsequent examinations</u> SFM after SFM: 0.83 FFDM after SFM: 1.00	screening populations outside of Norway." • "The NBCSP offers mammographic screening biennially, which differs from annual screening practices in some countries such as the United States." • "the use of consensus in double-reading practices in Norway is not standard in other population-based screening programs, including those in U.S. practice." • "We were not able to account for the effects of hormone replacement therapy in our screening population. Prior studies suggest that hormone therapy in postmenopausal women increases the risk of mammograms with false-positive results (29,30). The declining use of hormone replacement therapy during the past decade could be a confounding variable	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				(p: <0.05 compared to SFM after SFM) FFDM after FFDM: 0.91 **Screen-Detected Cancer for Invasive Breast Cancer (per 1,000 examinations)** • Overall SFM: 0.47 FFDM: 0.42 p: <0.001 • Baseline examinations SFM: 5.53 FFDM: 4.85 p: <0.05 • Subsequent examinations SFM after SFM: 4.21 FFDM after SFM: 4.01 FFDM after FFDM: 4.02 **PPV after Recall Examination after Mammography (%)** • Baseline examinations SFM: 12.9 FFDM: 10.0	in our observed decreased recall rates during the transition from SFM to FFDM.”	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				p: <0.05 • Subsequent examinations SFM after SFM: 19.3 FFDM after SFM: 21.63 (p: <0.05 compared to SFM after SFM) FFDM after FFDM: 22.73 (p: <0.05 compared to SFM after SFM) **Adjusted Incidence Rate Ratio (IRR) for Screening-detected Breast Cancer (95% CI)** • SFM after SFM 1.00 • FFDM after SFM 1.05 (0.98, 1.14) • FFDM after FFDM. 1.05 (0.96, 1.14) **Adjusted IRR for Screening-detected Invasive breast cancer (95% CI)** • SFM after SFM 1.00 • FFDM after SFM 0.99 (0.91, 1.07) • FFDM after FFDM		

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				1.00 (0.90, 1.10) **Adjusted IRR for Screening-detected DCIS (95% CI)** • SFM after SFM 1.00 • FFDM after SFM 1.43 (1.20, 1.71) • FFDM after FFDM 1.32 (1.07, 1.64) * IRR adjusted for screening modality, period, and age		
Van Ongeval, 2010 [Belgium]	• N/A	• Program: Decentralized screening program • Study Period: June 2001 – June 2009 • Target Age: 50 – 69 years • Screening Frequency: biennial • Sample Size (# of women): **Comparison Groups** DM population: 11,355 SFM 1st control population: 23,325	• Factor of study: Digital mammography (DM) vs. Screen-film mammography (SFM) • Other potential influencing factors: **Reading Approach** Local radiologist is the first reader. Second reader is from a CBU. Independent double reading with use of third reader "if only one of both readers	**Compared to First Control (SFM) Population* **Recall Rate (%)** • Initial DM: 2.64 SFM: 2.40 p: 0.43 • Subsequent DM: 1.20 SFM: 1.58 p: 0.03 **Cancer Detection Rate (%)** • Initial DM: 0.63 SFM: 0.60 p: 0.80	**Author Reported Conclusions** • "This is the first report on the results of a decentralized screening organization where DM is implemented successfully, with high CDR and without increase of the recall rate." **Author Reported Limitations** • "There are some shortcomings to the study. The first one is that, until now, there has been no exact	• "In decentralized screening programs, images are acquired in regional screening units (RSU) and these images are sent to a central breast unit (CBU). All screening activities, including the independent second and third reading, are performed in the CBU" • First control population: time period of SFM use among three RSU

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			scored a 3 or higher” **Reading Volume** Second readers: at least 5,000 examinations/year **Quality Assurance and Accreditation** “A national quality assurance manual, based on the European Guidelines, has been developed for SFM as well as for DM.” “For DM, a type testing procedure has been introduced: only systems that successfully passed the type test protocol can be presented for an acceptance test [10]. The type test protocol copies the European Guidelines for physical-technical quality assurance	• Subsequent DM: 0.57 SFM: 0.72 p: 0.22 **DCIS (%)** • Initial DM: 0.07 SFM: 0.16 p: 0.02 **PPV (%)** • Initial DM: 24.05 SFM: 24.86 p: 0.88 • Subsequent DM: 48.00 SFM: 45.93 p: 0.75 ** Compared to Second Control (SFM) Population* **Recall Rate (%)** • Initial DM: 2.64 SFM: 2.75 p: 0.70 • Subsequent DM: 1.20 SFM: 1.14 p: 0.66	registration of interval cancers by the Flemish government. However, a study in our screening region [16] showed that the percentage of interval cancers is compatible with the European Guidelines.” • “A second shortcoming is related with the relatively small number of centers involved.” • “The variation of expertise in reading screening mammography of the first readers in a decentralized screening organization is high, but this had no impact on the performance parameters of the screening organization.”	that were the first to change to DM • Second control population: “indicators of SFM of 47 mammographic units …cooperating with the CBU from 2001 till June 2008”

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			(QA) and adds a radiological evaluation on 25 examinations to assess stability of processing and the global appearance of the images [11, 12].” “systems in all the participating mammography units have to pass an acceptance test with all the criteria of the European protocol and a set of ten images is used to verify the transition of the images from the first center to the CBU center and to verify the global image quality and visualization on the work station in the center for second reading.” “Accreditation of the	**Cancer Detection Rate (%)** • Initial DM: 0.63 SFM: 0.69 p: 0.68 • Subsequent DM: 0.57 SFM: 0.47 p: 0.19 **PPV (%)** • Initial DM: 24.05 SFM: 25.29 p: 0.80 • Subsequent DM: 48.00 SFM: 41.29 p: 0.18		

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			unit requires that daily, weekly and (half-)yearly physical-technical quality control of the mammography system and the viewing station is being performed." **Radiologist Training** "…4 h of theory and 4 h of practice in reading DM mammograms (a total of 100 patient cases) and pass a reading examination on DM images" **Performance Feedback** "Since 2001, a continuous evaluation is done for all first and second readers by means of a number of parameters, including the individual recall rates, the number			

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			of readings and the discordance between the first and second reader. Several training courses and meetings are organized to improve possible poor results.” **Mammographic Views** Two views **Radiologist Experience** “Before 2001 there were some pilot projects of local organized breast cancer screening ,… one of the second readers was involved from the start in 1984 and therefore had substantial experience in screening. As the first readers had no experience with screening			

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			mammography, a slight bias on the results of the 2001 period due to the learning curve of the first readers cannot be excluded.” **Prior Mammograms** Prior mammographic examinations available during subsequent rounds			
van Luijt, 2013 [Netherlands]	• N/A	• Program: Dutch national breast cancer screening programme • Study period: 2004 – 2010 • Target age: 50 – 75 years • Screening frequency: biennially • Sample size (# of screens) = 6,007,582 **Technology & Study Group** DM: 1,452,508 SFM: 1,460,344 SFM only: 3,094,730	• Factor of study: Digital mammography (DM) vs. Screen-film mammography (SFM) • Other potential influencing Factors: **Radiologist Training** Training in the National Training and Reference Center (NETC) for 8 days **Reading Volume (Range)** Mean of 13,000 screens/year (3,000 to 60,000)	**Recall Rate (%)** • DM: 2.0 • SFM: 1.6 (p: <0.001) • SFM Only: 1.6 (p: <0.001) **Detection Rate (per 1,000 screens)** • DM: 5.9 • SFM: 5.1 (p: <0.001) • SFM Only: 5.0 (p: <0.001) **PPV (%)** • DM: 31.2 • SFM: 34.4 (p:<0.001) • SFM Only: 34.2 (p: <0.001)	**Author Reported Conclusions** • “In accordance to previous, smaller, studies, we can confirm that DM has a higher detection rate compared to SFM, at the cost of a higher recall rate and lower PPV. More DCIS and a higher fraction of very small tumours were detected with DM, which has positive consequences for the stage shift as a result of mass screening.”	• DM-group: “DM read by a reading unit reading both SFM and DM” • SFM-group: “SFM read by a reading unit reading both SFM and DM” • SFM only-group: “SFM read by a reading unit reading only SFM” • Article also reported performance measures by age groups, first screens, and timely subsequent screens • Article also provided figures of

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Mammographic Views Two view mammography used in 93% of screens in 2010		**Author Reported Limitations** • "We found that PPV is slightly lower in DM, but steadily increases over time. This must be due to a learning curve with better understanding of the findings on DM." • "The sharpest increase was found in the proportion of DCIS detected. This immediately triggers the concern of overdiagnosis, as often raised by those opposing screening." • "The increase in the number of DCIS detected is rather steep. Possibly this is due to a first pass effect. This means that outcomes will be most strongly affected directly after the introduction of a new technology or method, comparable to a prevalence screen. The number of DCIS might stabilise at a lower	recall rate, detection rate, and positive predictive value at screening over time by groups of study • Article also provided figure of the proportion of ductal carcinoma in situ (DCIS) by age groups, first screens, and timely subsequent screens • Article also provided range of variation in regions for performance measures by technology and age groups • Article also provided table with performance measures from other European research

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					level in the upcoming years." • "A wide range of variance exists when looking at regions separately." • "In 2010 93% of all participants were examined using a two view examination. This change in policy may also have influenced the performance rates of the screening programme."	
Sala, 2009	• More recent publication from this research group is available					
Juel, 2010	• Non-conventional FFDM – photon counting detector					
Lipasti, 2010	• Only women 50 to 59 years of age included					
Feeley, 2011	• Another study based on data from the Irish National Screening Program (Hambly et al. 2009) includes a larger sample					
Perry, 2011	• Patients allocated to screening					

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
	modality – not a real- world scenario					
Sala, 2011	- More recent publication from this research group is available					
van Ravesteyn, 2012	- No comparison of two technologies					
Chiarelli, 2013	- Concurrent cohorts [preference is given to studies exploring the effects of changes in technology over time]					
Comas, 2014	- Costs of switching to digital mammography					
Dabbous, 2017	- US study, probably concurrent cohorts					
Computer-Aided Detection (CAD)						
Bargallo, 2014 [Barcelona]	- N/A	- Program: Population-based breast cancer screening program (Sants-Montjuic, Les Corts, and Eixample Esquerre) - Study period: 2004 – 2012 DR+Arb: 2004-2010 SR+CAD: 2010-2012 - Target age: 50-69 years	- Factor of study: double reading with arbitration (DR+Arb) vs. single reading with CAD (SR+CAD) by experienced radiologist - Other potential influencing factors: **Reading Approach:**	**Recall Rate (%)** - SR+CAD: 7.02 - DR+Arb: 3.94 **Cancer Detection Rate (%)** - SR+CAD: 6.10 - DR+Arb: 5.25	**Author Reported Conclusions** - "The cancer detection rate of the screening program improved using a single reading protocol by experienced radiologists assisted by CAD, at the cost of a moderate increase	- Historical control study performed

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Screening frequency: every 2 years • Sample size (# of screens) DR+Arb: 47,462 SR+CAD: 21,321	In DR+Arb cohort, independent blinded double reading performed **Radiologist Experience:** DR+Arb: Arbitration conducted by third radiologist with most experience. SR+CAD: Radiologists with good historical performance based on cancer detection and recall rates were selected. **Technology** Full-field digital mammography (FFDM) with **Mammographic Views** Two views of each breast (craniocaudal and mediolateral oblique) **Audit and Feedback**	**Positive Predictive Value of Recall (%)** • SR+CAD: 8.69 • DR+Arb: 13.32	of the recall rate mainly related to the lack of arbitration." • "Radiologists specialized in breast imaging performed better than general radiologists." • "CAD did not detect any cancer not perceived previously by the radiologist." **Author Reported Limitations** • "there was not a unified regional tumor registry to establish the sensitivity and specificity. Consequently, we used the CDR and the RR as estimations of these variables." • "…although none of the readers admitted that CAD had modified their intention to recall in cancers, it is possible that CAD had in fact influenced the radiologists. To clarify this point, it would have been necessary	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Program includes clinical audit and readers receive yearly feedback **Reading Volume** 2700 screening mammograms/year (average)		to not allow the radiologists to check the CAD marks until they had issued a report. This would have clearly separated the opinion of the radiologist with and without CAD.”	
Sanchez Gomez, 2011 [Spain]	• N/A	• Program: “Population-based breast cancer screening program organized by the local government of La Rioja” • Study period: 3-year study • Screening frequency: biennially • Sample size (# of asymptomatic women) = 21,855 • Age range: 45-65 years	• Factor of study: Single reading vs. Single reading with CAD (or pre- vs. post-CAD interpretation) • Other potential influencing factors: **Technology** Screen-film mammography (SFM) **Mammographic views** Two views of each breast (craniocaudal and mediolateral oblique) **Radiologists’ Experience**	**Recall Rate (%)** • Pre-CAD: 7.2 • Post-CAD: 7.6 (increase not statistically significant) **Detection Rate (per 1,000 women)** • Pre-CAD: 4.3‰ • Post-CAD: 4.4‰ (p= <0.005) • “...detection rate was 4.3 carcinomas per 1000 women studied (10.5% were multiple lesions). CAD prompted a change of attitude in only one of the cases. Therefore, CAD supposed an increase	**Author Reported Conclusions** • “CAD supposed a significant increase in detection, without modifications in recall rates and PPV of biopsy. However, better results could have been achieved if radiologists had considered actionable those cases marked by CAD but initially misinterpreted.” **Author Reported Limitations** • “When we analyzed delayed diagnosis due to radiologist’s false negatives results,	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Specialized breast radiologists: 2-10 years of experience in mammography **Technologist Experience** >8 years of mammography experience **Radiologists' Training** General radiologists: 3-9 months of training in reading of screening mammograms **Reading Approach** Single reading	of 0.1‰ in detection rate and 1% in the total number of cases ($p < 0.005$).” **PPV (%)** • Radiologist: 6.4 • CAD: 0.46 • Both: 6.1 • “PPV of percutaneous biopsy was unchanged by CAD (20.23% pre- and post-CAD interpretation).”	61.5% occurred in the group of general radiologists, while 38.5% were missed by specialized breast radiologists. Although these results are not statistically significant due to the low number of missed cases, in our opinion this difference can be related to poorer interpretation skills among general radiologists, who are more prone to misinterpret, either false or true positive CAD-marks, in comparison to specialized radiologists. We consider this fact as a bias in our study, as the inclusion of general radiologists is not a standard and it can have potential effects in increasing recall rates and decreasing detection rates if CAD is used as a complementary detection tool by less experienced radiologists.”	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					- "In our study, the number of missed cancers was higher than that in Helvie's (13 and 2 misinterpreted cases, respectively). This fact could be related to the inclusion of general radiologists in our study, with poorer results as referred above."	
Fenton, 2011 [USA]	- N/A	- Program: Breast Cancer Surveillance Consortium (BCSC) - Study period: January 1, 1998 - December 31, 2006 - Sample size (# of screens) = > 1.6 million - Sample size (# of radiologists) = 793 - Age (years): ≥ 40 years	- Factor of study: Screen-film mammography (SFM) vs. SFM with computer aided detection (CAD) - Other potential influencing factors: **Time to Adapt** "We excluded mammograms interpreted during the initial 3 months of CAD use at each facility because radiologists may have been adapting to the technology during this period."	**Unadjusted Recall Rate (%; 95% CI)** - Never implemented CAD 1998-2002: 9.3 (9.2, 9.3) 2003-2006: 9.1 (9.0, 9.2) p= 0.005 - Implemented CAD Before CAD: 8.4 (8.3, 8.5) After CAD: 8.9 (8.8, 9.0) p= <0.001 **Unadjusted All Breast Cancer Detection Rate (per 1,000 mammograms; 95% CI)**	**Author Reported Conclusions** - "Among a large sample of US mammography facilities, CAD was associated with statistically significantly decreased specificity and PPV1. CAD was not associated with improved sensitivity for invasive breast cancer, increased rates of breast cancer detection, or more favorable stage or size of invasive breast cancers. CAD is now applied to the large majority of screening mammograms in the United States with	- CAD was implemented in 25 of 90 BSCS facilities - Article also provided unadjusted positive predictive value by months since CAD implementation

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				• Never implemented CAD 1998-2002: 4.2 (4.0, 4.3) 2003-2006: 4.0 (3.8, 4.3) p= 0.30 • Implemented CAD Before CAD: 3.6 (3.4, 3.8) After CAD: 3.2 (3.0, 3.5) p= 0.01 **Unadjusted Invasive Breast Cancer Detection Rate (per 1,000 mammograms; 95% CI)** • Never implemented CAD 1998-2002: 3.3 (3.2, 3.5) 2003-2006: 3.2 (3.0, 3.4) p= 0.27 • Implemented CAD Before CAD: 2.8 (2.7, 3.0) After CAD: 2.3 (2.1, 2.5) p= <0.001	annual direct Medicare costs exceeding $30 million (2). As currently implemented in US practice, CAD appears to increase a woman's risk of being recalled for further testing after screening mammography while yielding equivocal health benefits" **Author Reported Limitations** • "A limitation of this study is the absence of digital mammography data. Whereas CAD algorithms perform a similar alerting function in the screen-film and digital environments, screen-film mammograms must be digitized before CAD analysis, and digitization may introduce noise and adversely affect performance." • "Because prior research suggests	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				Unadjusted Ductal Carcinoma In Situ (DCIS) Detection Rate (per 1,000 mammograms; 95% CI) • Never implemented CAD 1998-2002: 0.9 (0.8, 0.9) 2003-2006: 0.9 (0.8, 1.0) p= 0.91 • Implemented CAD Before CAD: 0.8 (0.7, 0.9) After CAD: 0.9 (0.8, 1.0) p=0.13 **Unadjusted DCIS Detection Rate (%)** • Never implemented CAD 1998-2002: 18.9 2003-2006: 19.2 p= 0.80 • Implemented CAD Before CAD: 20.0 After CAD: 24.9 p= 0.003	that facilities apply CAD on nearly all mammograms after implementation (10), these analyses assumed that all mammograms were interpreted with CAD after implementation—another limitation of this study. To the extent that facilities did not use CAD on all mammograms, results may be biased toward the null." • "As the analyses account for salient patient factors, unmeasured radiologist or facility characteristics may affect results." • "Although the number of women with breast cancers diagnosed after CAD implementation (>1000 cancers) is greater than that observed in previous samples, larger samples may be needed to detect small increases in	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				Unadjusted Positive Predictive Value (%; 95% CI) • <u>Never implemented CAD</u> 1998-2002: 4.5 (4.3, 4.7) 2003-2006: 4.4 (4.2, 4.7) p= 0.63 • <u>Implemented CAD</u> Before CAD: 4.3 (4.1, 4.5) After CAD: 3.6 (3.4, 3.9) p= <0.001 **Adjusted OR for PPV$_1$ among CAD Use vs. Non-CAD Use (95% CI)** • OR: 0.89 (0.80, 0.99) • p= 0.03 **Adjusted OR for Detection of Any Breast Cancer among CAD Use vs. Non-CAD Use (95% CI)** • OR: 1.01 (0.92, 1.12) • p= 0.79 **Adjusted OR for Detection of Invasive**	sensitivity or cancer detection with CAD." • "Finally, another limitation of this study is the lack of data on the CAD products that each facility used, so the potentially distinct impacts of different products could not be investigated."	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				Breast Cancer among CAD Use vs. Non-CAD Use (95% CI) • OR: 0.97 (0.86, 1.08) • p= 0.54 **Adjusted OR for Detection of DCIS among CAD Use vs. Non-CAD Use (95% CI)** • OR: 1.16 (0.95, 1.42) • p= 0.14 * "Odds ratios were adjusted for mammography registry, patient age (40–44, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, and ≥75 years), breast density (almost entirely fat, scattered fibroglandular tissue, and heterogeneously and extremely dense), time since prior mammography (no prior mammogram, 9–15 months, 16–20 months, 21–27 months, and ≥28 months), current hormone replacement		

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				therapy, and year of examination (1998–2002 or 2003–2006)."		
Lehman, 2016 [USA]	- N/A	- Program: Breast Cancer Surveillance Consortium (BCSC) - Study period: January 1, 2003 - December 31, 2009 - Sample size (# of screens) = 625,625 **Technology** With CAD: 495,818 Without CAD: 129,807 - Sample size (# of radiologists) = 271 - Age (years): 40 – 89 years	- Factor of study: digital mammography (DM) vs. DM with computer-aided detection (CAD) - Other potential influencing factors: **Account for Potential Learning Curve** "To allow for the possibility that performance improved after the first year of CAD use by a radiologist, and to account for any possible learning curve, we excluded the first year of mammographic interpretations with CAD for individual radiologists and found no differences for any of our performance	**Recall Rate Per 100 Exams (Mean; 95% CI)** - CAD: 8.7 (8.1, 9.4) - No CAD: 9.1 (8.4, 9.8) **Adjusted OR for Recall Rate among CAD vs. No CAD (95% CI)** - OR: 0.96 (0.89, 1.04) - p= 0.35 **Total Cancers Detected Per 1,000 Exams (Mean; 95% CI)** - CAD: 4.06 (3.8, 4.4) - No CAD: 4.10 (3.6, 4.6) - (no significant difference) **Adjusted OR for Total Cancers Detected among CAD vs. No CAD (95% CI)** - OR: 0.99 (0.84, 1.15) - p= 0.86 **Invasive Cancers Detected Per 1,000 Exams (Means; 95% CI)**	**Author Reported Conclusions** - "We found no evidence that CAD applied to digital mammography in U.S. community practice improves screening mammography performance on any performance measure or in any subgroup of women. In fact, mammography sensitivity was decreased in the subset of radiologists who interpreted mammograms with and without CAD. This study builds on prior studies (18–19) by demonstrating that radiologists' early learning curve and patient characteristics do not account for the lack of benefit from CAD."	- Article also provided cancer detection results for CAD/No CAD by age, BI-RADS breast density, menopausal status, and time since last mammogram

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			measurements (data not shown)."	- CAD: 2.91 (2.7, 3.1) - No CAD: 3.05 (2.7, 3.5) - (no significant difference) **Adjusted OR for Invasive Cancers Detected among CAD vs. No CAD (95% CI)** - OR: 0.92 (0.77, 1.08) - p= 0.30 **Ductal Carcinoma In Situ (DCIS) Detected Per 1,000 Exams (Means; 95% CI)** - CAD: 1.19 (1.0, 1.3) - No CAD: 0.95 (0.7, 1.2) - p= <0.03 **Adjusted OR for DCIS Detected among CAD vs. No CAD (95% CI)** - OR: 1.39 (1.03, 1.87) - p= 0.031 * "Odds ratio for CAD vs. No CAD adjusted for site, age group, race/ethnicity, time since prior	**Author Reported Limitations** - "Given the observational methods of our study, we could not compare mammography performance among women who had their mammograms interpreted both with and without CAD. It is possible that CAD was used preferentially in women whose mammograms were more challenging." - "We also were not able to control for radiologist characteristics, such as experience, and thus compared performance with and without CAD in the same radiologists, to address across-radiologist variability."	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				mammogram and calendar year of the exam using mixed effects model with random effect for exam reader and varying with CAD use.”		
Gromet, 2008 [USA]	- N/A	- Program: community-based mammography program in Charlotte, NC - Study period: January 1, 2001-December 31, 2005 - Sample size (# of screens) = 231,221 - Sample size (# of radiologists) = 9 - Mean age (years): double-read patients: 53.8 (SD, 11.5); single-read CAD patients:53.5 (SD, 11.1)	- Factor of Study: Comparison 1 Single reading with computer-aided detection (CAD) vs. Double reading Comparison 2 Single reading with CAD vs. First reader in a double-reading program - Other potential influencing factors: **Reading Approach** Batch reading; double reading until 2003; conversion to single reading with CAD during 2003. Double reading approach: cases classified as negative by the	**Recall Rate (%)** - Single reading with CAD vs. Double reading Single reading + CAD: 10.6 Double reading: 11.9 p= <0.0001 - Single reading with CAD vs. First reader Single reading + CAD: 10.6 First reader: 10.2 p= <0.0001 **Detection Rate (per 1,000 patients)** - Single reading with CAD vs. Double reading Single reading + CAD: 4.2 Double reading: 4.46	**Author Reported Conclusions** - Single reading with CAD vs. double reading “Single reading with CAD showed no statistically significant difference from double reading in sensitivity, cancer detection rate, or PPV1. However, the recall rate was lower with CAD (10.6%) than with double reading (11.9%). The 1.3% difference was statistically significant (p < 0.0001).” - Single reading with CAD vs. first reader “Compared with the first reader	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			first reader and positive by the second reader were resolved by a different subspecialist reader who determined the final reading. **Readers' Training** radiologists; first readers were specialized mammographers; second reading was performed by a general radiologist with certification in mammography who did not specialize in the area. Experience: 1-24 years (mean 15 years). The only radiologist with <5 years of experience joined directly after fellowship training. Annual volume from 4,459 to 15,281 readings. **Prior Mammograms:**	p=0.347 • <u>Single reading with CAD vs. First reader</u> Single reading + CAD: 4.2 First reader: 4.12 p= 0.761 **PPV$_1$ (%)** • <u>Single reading with CAD vs. Double reading</u> Single reading + CAD: 3.9 Double reading: 3.7 p= 0.371 • <u>Single reading with CAD vs. First reader</u> Single reading + CAD: 3.9 First reader: 4.1 p= 0.662 **Sensitivity (%)** • <u>Single reading with CAD vs. Double reading</u> Single reading + CAD: 90.4 Double reading: 88.0 p= 0.205 • <u>Single reading with CAD vs. First reader</u>	performance in the double-reading program, single reading with CAD resulted in a significantly increased sensitivity (90.4% vs 81.4%, respectively; p < 0.0001) at a cost of a small increase in the recall rate (10.6% vs 10.2%, p < 0.0001). There was no statistically significant difference in PPV1 or cancer detection rate." • Overall: "…both double reading and CAD are effective methods to increase the sensitivity of screening mammography for experienced mammogram readers. In our study, the second reader increased sensitivity 6.6%, from 81.4% to 88.0%; the recall rate rose from 10.2% to 11.9%. Single reading	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			If available, preference was given to a 3-year old prior examination; additional prior films were available on request **Technology** Screen-film	Single reading + CAD: 90.4 First reader: 81.4 p= <0.0001	enhanced by CAD review yielded a higher sensitivity of 90.4%, with a smaller increase in the recall rate from 10.2% to 10.6%. With manpower and cost constraints limiting the use of double reading in the United States, CAD appears to be an effective alternative that provides similar, and potentially greater, benefits." **Author Reported Limitations** • possible effect of improved radiologists' skills over time on performance [performance was before and after CAD implementation was compared]	
James and Cornford, 2009 [UK]	• N/A	• Study: Retrospective study • Study period: July 2003 – April 2004	• Factor of Study: Computer-aided detection (CAD) vs. third reader in double reading with arbitration	**Recall Following Arbitration** • CAD increases discordant double-reading cases recalled from 47% to 68%	**Author Reported Conclusions** • "The present study has shown that the main effect of CAD acting as an arbitrator	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Sample size (# of women) = 16,629	• Other potential influencing factors: **Reading Approach** unblinded double reading with arbitration; the third reader had knowledge of the first two readers' opinions **Readers' Training** Five consultant radiologists (5 to 18 years of experience), one research fellow, one radiographer (5 years of experience); the research fellow and the radiographer did not act as arbitrators **Mammographic Views** Two views **Technology** Film (mammograms of the arbitration cases were	• Additional 50 women recalled with CAD could possibly result in a relative increase in the overall recall rate by 10% (3.1% to 3.4%) **Recall for Cancer Following Arbitration** • 3rd Reader Arbitration 83% (15/18) of the cancers recalled • CAD Arbitration 89% (16/18) of the cancers recalled • One additional cancer detected with CAD **Normal Women Recalled Following Arbitration** • 3rd Reader Arbitration 43.9% (94/214) normal cases recalled • CAD Arbitration 66.8% (143/214) normal cases recalled • Significant increase in normal women recalled for assessment (P<0.001)	of discordant double-reading opinions is to increase the recall rate, significantly above what is found when arbitration is performed by an independent third reader. Numbers of cancers detected were broadly similar with one additional cancer case recalled when CAD acted as the arbitrator. The use of an independent third reader to arbitrate discordant double-reading opinions remains the best method of maintaining high cancer detection whilst keeping recall rates low. It may be that using CAD as an arbitrator may be an option to deal with discordant double-reading opinions when no other method of consensus or arbitration is available.” **Author Reported Limitations**	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			digitalized to be analyzed by a CAD system)	with CAD	• "The study is retrospective and so can only give an indication as to the potential effect of a CAD system acting as an arbiter of discordant double-reading opinions." • "...due to the small number of cancers, the lack of statistical power makes meaningful statistical analysis impossible."	
			Tomosynthesis			
Friedewald, 2014 [USA]	• N/A	• Program: retrospective analysis of screening performance metrics from 13 academic and non-academic breast centers • Study period: March 2010-December 31, 2012 Period 1: one year before tomosynthesis implementation (start dates March 2010-October 2011) Period 2: DM+tomosynthesis examinations after initiation of tomosynthesis screening from March-	• Factor of Study: digital mammography (DM) with tomosynthesis vs. DM alone • Other potential influencing factors: **Reading Approach:** not reported **Readers' Training:** radiologists (N=139 in all sites) **Prior Mammograms:** not reported	• **Recall Rate (per 1000)** DM alone: Actual: 106 Model estimate: 107 (95% CI: 89, 124) DM+tomosynthesis: Actual: 89 Model estimate: 91 (95% CI: 73, 108) Change DM+tomosynthesis vs. DM alone: -16 (95% CI: -18, -14). P<0.001	**Author Reported Conclusions** • The addition of tomosynthesis to digital mammography was associated with a decrease in recall rate and an increase in cancer detection rate." **Author Reported Limitations** • "...lack of a randomized trial design, in which 2 cohorts are concurrently enrolled and screened, introduces the	• Model estimates adjustment for site as a random effect and time period as a fixed effect

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		October 2012 through December 31, 2012 • Target age: not reported • Screening frequency: • Sample size (# of screens) = 454,850 DM alone: 281,187 DM+tomosynthesis: 173,663 • Mean age (years): DM alone: 57.0 DM+tomosynthesis:56.2	**Technology:** factor of study	11 of 13 sites observed a decrease in recall rates with DM+tomosynthesis; 2 sites observed an increase of 18 per 1000 examinations • **Cancer Detection Rate per 1000** DM alone: Actual: 4.3 Model estimate: 4.2 (95% CI: 3.8, 4.7) DM+tomosynthesis: Actual: 5.5 Model estimate: 5.4 (95% CI: 4.9, 6.0) Change DM+tomosynthesis vs. DM alone: 1.2 (95% CI: 0.8, 1.6) 12 of 13 sites observed an increase in cancer detection rates • **Invasive Cancer Detection Rate (per 1000)**	possibility that results were not purely due to the addition of tomosynthesis. … However, there were no differences in mean age between the 2 periods, and the use of the same sites in both periods was intended to provide comparable populations in the 2 cohorts. We would not expect the risk profile at any given site to change meaningfully between the 2 periods, and our statistical models adjusting for site effects were consistent with the unadjusted results." • "The fact that sites converted incrementally to tomosynthesis further introduces the possibility of selection bias. However, sensitivity analysis including the concurrent digital mammograms in the tomosynthesis period	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				DM alone: Actual: 2.9 Model estimate: 2.9 (95% CI: 2.5, 3.2) DM+tomosynthesis: Actual: 4.1 Model estimate: 4.1 (95% CI: 3.7, 4.5) Change DM+tomosynthesis vs. DM alone: 1.2 (95% CI: 0.8, 1.6). P<0.001 12 of 13 sites observed an increase in invasive cancer detection rates • **PPV (%)** DM alone: Actual: 4.1 Model estimate: 4.3 (95% CI: 3.4, 5.3) DM+tomosynthesis: Actual: 6.1 Model estimate: 6.4 (95% CI: 5.4, 7.4)	suggested that selection bias alone could not account for the significant performance gains.” • “…only population-level (rather than patient-level) statistics were available from each site. Therefore, we were not able to evaluate the number of repeat examinations and, as a consequence, avoided statistical assumptions of independent observations.” • “While implementation of tomosynthesis in our study was associated with a reduction in recall rate from screening, follow-up data were not available that would allow evaluation of false-negative result rates. The study did not assess clinical outcomes, so whether the increase in cancer detection	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				Overall change DM+tomosynthesis vs. DM alone: 2.1 (95% CI: 1.7, 2.5). P<0.001	rates is of benefit is not known."	
Rafferty, 2017 [USA]	- N/A	- Same multi-center study as described by Friedewald, 2014 (above)	- See Friedewald, 2014 (above)	- "Addition of tomosynthesis to digital mammography produced significant reductions in recall rates for all age groups and significant increases in cancer detection rates for women 40–69. Largest recall rate reduction with tomosynthesis was for women 40–49, decreasing from 137 (95% CI 117–156) to 115 (95% CI 95–135); difference, -22 (95% CI -26 to -18; P<.001). Simultaneous increase in invasive cancer detection rate for women 40–49 from 1.6 (95% CI 1.2–1.9) to 2.7 (95% CI 2.2–3.1) with tomosynthesis (difference, 1.1; 95% CI 0.6–1.6; P<.001) was observed."	- "Addition of tomosynthesis to digital mammography increased invasive cancer detection rates for women 40–69 and decreased recall rates for all age groups with largest performance gains seen in women 40–49. The similar performance seen with tomosynthesis screening for women in their 40s compared to digital mammography for women in their 50s argues strongly for commencement of mammography screening at age 40 using tomosynthesis."	- Rafferty et al. (2017) analyzed the same data as Friedewald et al. (2014) with the aim to determine the effect of patients' age on the performance of tomosynthesis+DM vs. DM alone

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
Giess, 2017 [USA]	• N/A	• Program: an academic medical center • Study period: October 2012 – May 2015 • Target age: not reported • Screening frequency: annual (see "Discussion" section, last paragraph on page 933) • Sample size (# of screens) = 37,338 FFDM: 16,264 FFDM+DBT: 21,074 • Mean age (years): 54.8±10.3	• Factor of Study: 2D full field digital mammography (FFDM) + digital breast tomosynthesis (DBT) vs. FFDM alone • Other potential influencing factors: **Reading Approach:** not reported **Readers' Training:** 15 breast imaging specialists and 9 general radiologists. Eighteen radiologists had >10 years of experience. All radiologists completed an 8-hour training program in DBT technology. **Prior Mammograms**: available for 87.8% of examinations **Technology:** factor of study	• **Recall Rate (%)** Overall FFDM: 10.3 FFDM+DBT: 10.7 P=0.26 Baseline examinations FFDM: 21.9 FFDM+DBT: 16.4 P<0.001 • **Cancer Detection Rate (per 1000)** Overall FFDM: 1.8 FFGM+DBT: 3.8 P=0.005 Baseline examinations FFDM: 1.45 FFGM+DBT: 3.3 P=0.32 • **PPV1 (%)** Overall FFDM: 1.8 FFGM+DBT: 3.6 P=0.006 Baseline examinations FFDM: 0.6 FFGM+DBT: 1.8 P=0.22	**Author Reported Conclusions** • "…the cancer detection rate and PPV of recalled examinations were higher with DBT than with FFDM. We found no significant difference in the recall rate between the two imaging techniques in our overall cohort, though DBT resulted in a lower recall rate in patients being screened for the first time (baseline screening)." • "We found no significant difference in cancer detection rate and PPV1 in baseline examinations of patients who underwent DBT compared with FFDM, though the subgroup analysis was likely underpowered to reach statistical significance because of the small number	• In this article, FFDM+DBT group is referred to as DBT (see "Screening Protocol and Interpretation" section) • This study was conducted in a mixed screening environment when some patients underwent FFDM while others FFGM+DBT. There may be differences in the risk factors for breast cancer between patients undergoing different screening modalities. Propensity score matching was used to reduce the bias due to variables significantly associated with differential imaging technic selection (DBT vs. FFDM).

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					of cancers in this subgroup." • "In conclusion, we found that DBT improves breast cancer detection at screening mammography. Although the effects on overall recall rate may be lower with DBT than previously reported, the improvement in PPV1 suggests a reduction in unnecessary recalls." **Author Reported Limitations** • "There were relatively low cancer detection rates in both the FFDM and DBT subgroups. In applying propensity scoring to adjust for nonrandom assignment of patients to DBT, 158 cancer cases (71% of screen-detected cancers) were excluded. Thus, our data reflect relative differences in cancer detection rates between FFDM and	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					DBT, rather than absolute detection rates, but detecting relative differences in cancer detection rate was our prime objective.” • “Our patient population is a highly screened one, because our geographic region has a high preponderance of well-educated relatively affluent women. Annual screening in a normal-risk population will yield a lower cancer detection rate than biennial or less frequent screening.” • “DBT was preferentially offered to baseline screening studies in our practice, which could affect the cancer detection rate by detecting both prevalent and incident cancers. However, we did not identify any significant difference in cancer detection	

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					rate between DBT and FFDM baseline studies.”	
Greenberg, 2014 [USA]	• This study and Friedewald et al. 2014 (summarized above) were both conducted in the USA, compared DM+DBT and DM alone, and reached the same conclusions. However, Friedewald et al. 2014 included a larger sample.	•	•	•		•
Hogue, 2016 [Canada, Quebec]	• This is a conference abstract; full text not found	•	•	•		•
Houssami, 2017 [Italy-? Australia-?]	• Two technologies were compared under different reading approaches; unclear whether technology or reading approach determined the outcome	•	•	•		•
Lourenco, 2015 [USA]	• N/A	• Program: a dedicated breast imaging center • Study period: <u>DM</u>: from March 2011	• Factor of Study: digital mammography (DM) vs. digital breast tomosynthesis (DBT)	• **Recall Rate (%)** DM: 9.5 (95% CI: 8.8, 9.9) DBT: 6.4 (95% CI: 6.0, 6.8)	**Author Reported Conclusions** • “In summary, our study showed a decreased recall rate	• “…an observational, abrupt switch, nonmatched pre- and/or post-DBT design was

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		through February 2012 DBT: March 2012 through February 2013 • Target age: not reported • Screening frequency: not reported • Sample size (# of screens) = 12,577 (DM)+12,921 (DBT) • Mean age [women with BIRADS category 0 assessment]: 54.6 years ± 10.7 (DM); 55.3 years ±10.8 years (DBT). P=0.15	• Other potential influencing factors: **Reading Approach:** Batch reading with CAD. **Readers' Training:** six fellowship-trained breast radiologists with 4-16 years of experience. All radiologists completed the required 8-hour tomosynthesis training **Prior Mammograms:** • **Technology:** factor of study	P<0.00001 "The recall rate was lower with DM than with DBT for masses (8.9% vs 26.8%, respectively), distortions (0.6% vs 5.3%), and calcifications (13.4% vs 20.3%) (P<0.0001 for all). The recall rate was lower with DBT than with DM for asymmetries (13.3% vs 32.2%, respectively) and focal asymmetries (18.2% vs 32.2%) (P<0.0001 for both)." • **Cancer Detection Rate (per 1000)** DM: 5.4 DBT: 4.6 P=0.44 • **PPV for recall (%)** DM: 5.8 DBT: 7.2 P=0.219 [see "Results", last paragraph] • "At the time of additional imaging, fewer patients were	without a change in biopsy PPV or cancer detection rate after implementation of DBT, along with fewer recalls for asymmetries and more recalls for masses, calcifications, and areas of architectural distortion. More patients were evaluated with US only, and fewer required additional mammographic views only at the time of additional imaging following DBT screening." **Author Reported Limitations** • "One limitation of our study is its retrospective design, which prevents causal inference from being drawn of our results." • "Accurate assessment of false-negative findings is also not performed because many patients have not yet returned for subsequent imaging."	implemented whereby only DM was used for the 1st year and DBT was subsequently used for the 2nd year, without assignment or patient choice. Therefore, no systematic selection bias is anticipated. Physician staffing was unchanged during the study period."

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				evaluated with additional mammographic views (40.2% before DBT, 28.4% with DBT) and more were evaluated with US [ultrasound] only (2.6% before DBT, 28.3% with DBT)."	• "Differences in visualization of calcifications may also be related, at least in part, to differences in digital detectors between the two vendors used during this study."	
McCarthy, 2014 [USA]	• This study and Friedewald et al. 2014 (summarized above) were both conducted in the USA, compared DM+DBT and DM alone, and reached the same conclusions. However, Friedewald et al. 2014 included a larger sample.	•	•	•		•
McDonald, 2015 [USA]	• This study was conducted in the USA and included only data from baseline (first) screening	•	•	•		•
McDonald, 2016	• Full text is unavailable	•	•	•		•

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
Powell, 2017 [USA]	- This study and Friedewald et al. 2014 (summarized above) were both conducted in the USA, compared DM+DBT and DM alone, and reached the same conclusions. However, Friedewald et al. 2014 included a larger sample.	•	•	•		•
Procasco, 2016	- Full text is unavailable	•	•	•		•
Sharpe, 2016 [USA]	- N/A	- Program: an academic medical center - Study period: January 3, 2011 to March 15, 2014 - Target age: not reported - Screening frequency: not reported - Sample size (# of screens) = 80,149 (2D mammography); 5,703 (DBT) - Mean age (years): 55.68±9.74 (DBT); 57.62±10.89 (2D)	- Factor of Study: Digital breast tomosynthesis (DBT) vs. two-dimensional (2D) mammography - Other potential influencing factors: **Reading Approach:** not reported **Readers' Training:** 89.2% of the examinations were interpreted by ten board certified breast-subspecialized radiologists; seven were fellowship-	- **Recall Rate (%)** 2D mammography: 7.51 DBT: 6.10 P<0.0001 - **Recall rate for first (baseline) screening (%)** 2D mammography: 19.60 DBT: 9.88 P<0.0001 - **Recall rates stratified by breast density** Recall rates were lower with DBT than with 2D mammography for all breast density groups;	**Author Reported Conclusions** - "Implementing DBT into a U.S. breast cancer screening program significantly decreased the screening RR overall and for certain patient subgroups, while significantly increasing the CDR. These findings may encourage more widespread adoption and reimbursement of DBT and facilitate improved patient selection."	- It is unclear whether DBT was used in combination with 2D or on its own. - Patients were assigned to the first available machine. Fewer than 0.05% of patients specifically requested 2D of DBT screening. Each request was resolved on a case-by-case basis. - "Breast densities and ages of the patients in the 2D mammography and

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		•	trained in breast imaging with on average 15.6 years of experience in breast imaging. Three radiologists without fellowship had 27, 35 and 41 years of experience in breast imaging. Each radiologist interpreted >150 2D and DBT examinations during the study period. 10.8% of the examinations interpreted by 12 lower-volume general radiologists were excluded from RR variation analysis. All radiologists received >8 hours of tomosynthesis training before interpreting DBT examinations. **Prior Mammograms:** not reported	statistically significant differences for heterogeneously dense (7.33% vs 9.31%, P = .0048) and extremely dense (4.74% vs 6.54%, P = .0429) breasts. • **Recall rates stratified by age:** Recall rates were lower with DBT than with 2D mammography for all age groups; significant differences in 40–49-year old (8.66% vs 10.93%, P= .0075) and 60–69-year-old patients (3.66% vs 5.86%, P = .0006). • **Cancer Detection Rate (per 1000)** 2D mammography: 3.5 DBT: 5.4 P<0.0018 • **Invasive cancer detection rate (per 1000)** 2D mammography: 2.46 DBT: 2.81 P=0.61	**Author Reported Limitations** • "…far more 2D mammographic examinations than DBT examinations, and this could be considered a limitation of our study." • "…some subgroups had relatively small sample sizes, and this may have contributed to some differences not reaching statistical significance." • "The percentage of invasive cancers was lower with DBT than with digital mammography. The rate per 1000 was not statistically significantly different. It is not known which of the additional cancers detected by using DBT would have progressed to invasive malignancy and which would not have. For some, this could be considered a limitation of this investigation."	DBT groups were similar. Patients who underwent DBT were more likely to have a personal history of breast cancer, to have a family history of breast cancer, to have a personal history of a breast biopsy with a benign result, and to be undergoing a baseline examination."

1st Author, Date [Country]	Reasons for Exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			- **Technology:** factor of study	- **In situ cancer detection rate** 2D mammography: 1.04 DBT: 2.63 P>0.0006		

Table A11. Quality Assurance Practices

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
1st Author, Date *[Country]*	- *Specify*	- *Program/Study Name* - *Study period* - *Target age* - *Screening frequency* - *Sample size* - *Age of women*	- *Factor of study: Quality assurance practices* - *Other potential influencing factors: technology, screen readers' characteristics, reading approach, etc.*	- *Recall rate* - *False positive* - *Cancer detection rate* - *Positive predictive value*	*Conclusions* - *Author reported conclusions* *Limitations* - *Author reported limitations*	- *Comments (if any)*
Reading Volume						
Alberdi, 2011 [Spain]	- N/A	- Program: four Spanish population-based breast cancer screening programs - Study period: March 1990 – December 2006 - Target age: 45-69 years - Screening frequency: biannual - Sample size (# of screens) = 1,440,384; (# women) = 471,112; (# radiologists) = 72	- Factor of Study: reading volume (number of mammograms read in the previous 365 days - Other potential influencing factors: **Reading Approach:** Single reading **Reading Volume:** - Only years in which radiologists interpreted at least 500 mammograms were included in these analyses **Technology**: "analog or digital, the latter	- **Overall false positive (FP) OR (95% CI); multivariate analysis** Reading volume (# of mammograms read in the previous year) 0-499 (ref.): 500-1,999: 0.77 (0.73, 0.81) P<0.001 2,000-4,999: 0.71 (0.68, 0.75) P<0.001 5,000-9,999: 0.76 (0.72, 0.80) P<0.001	**Author Reported Conclusions** - "A decreasing tendency in the risk of overall false positive results was found as the reading volume in the previous year increased. Specific estimations of risk revealed a cut-off point above 10,000 readings in the previous year, with a lower limit of the confidence interval that did not overlap with any of the categories of less than 10,000 readings per year. The reduced risk of a false positive result with	- Article also reports on the effect of years of service (see Reader's characteristics) - In the author's view, the effect of reading volume on the rate of FP was greater than the effect of years of service. - Data on radiologists' experience (years of service and reading volume) were obtained from screening program databases (in

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Mean age (years): not reported	being considered only if performed and read in a digital format"	10,000-14,999: 0.62 (0.59, 0.65) P<0.001 >15,000: 0.59 (0.57, 0.62) P<0.001 • **FP leading to an invasive procedure OR (95% CI); multivariate analysis** Reading volume (# of mammograms read in the previous year) 0-499 (ref.): 500-1,999: 0.78 (0.66, 0.92) P=0.004 2,000-4,999: 0.78 (0.66, 0.92) P=0.003 5,000-9,999: 0.75 (0.64, 0.87) P<0.001 10,000-14,999: 0.56 (0.47, 0.65) P<0.001 >15,000: 0.60 (0.51, 0.70) P<0.001	greater reading volume was also observed with a similar magnitude for false-positives resulting in an invasive procedure but without a clearly differentiated cut-off point, as the confidence intervals of the OR overlapped between categories." **Author Reported Limitations** • "…radiologist experience outside the screening programme was not taken into account."	contrast to other studies that rely on self-reported data) • Cancer detection rates, PPV or sensitivity are not reported • Adjustment for the number of mammographic views, (one or two), mammogram type (analogue or digital), screen type (first or subsequent), period when the mammogram was performed (in 2-year intervals), and patient's age. The radiology unit where the mammogram was performed was included in the model as a random effect.
Barlow, 2004 [USA]	• N/A	• Program: mammography registries participating in the Breast Cancer	• Factor of Study: Reading volume • Other potential influencing factors:	• **Recall Rate (%)** No. of mammograms interpreted in the past year: ≤1000: 7.6 1001-2000: 11.1	**Author Reported Conclusions** • "In our study, radiologists with higher volumes did show	• Article also reports data on radiologists' characteristics (age, gender, experience,

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		Surveillance Consortium (BCSC) • Study period: January 1996 – December 2001 • Target age: ≥40 years • Screening frequency: not reported. To be included, a mammogram had to occur ≥9 months after any proceeding breast imaging to avoid misclassifying a diagnostic examination as screening • Sample size (# of screens) = 469,512; (# women) = 308,634; (# radiologists) = 124 • Mean age (years): not reported. Distribution by age categories (absolute numbers) is reported in table 1.	**Reading Approach:** Not reported **Patient Characteristics Considered:** Breast density, previous mammography, age, mammography registry **Technology:** not reported	>2000: 10.4 % of mammograms interpreted in the past year that were screening mammograms: <50: 9.5 51-75: 10.0 76-100: 10.7 • **Recall OR (95% CI); adjusted for patient's characteristics** No. of mammograms interpreted in the past year: ≤1000: 1.00 (ref.) 1001-2000: 1.51 (1.12, 1.82) >2000: 1.29 (0.97, 1.35) P=0.002 % of mammograms interpreted in the past year that were screening mammograms: <50: 1.00 (ref.) 51-75: 1.06 (0.73, 1.55) 76-100: 0.99 (0.68, 1.44) P=0.77 • **Sensitivity (95% CI); adjusted for patients' characteristics** No. of mammograms interpreted in the past year: ≤1000: 1.00 (ref.)	higher recall rates and higher sensitivity but lower specificity." • "Increasing volume requirements is unlikely to improve overall mammography performance. " • "…unless there is adequate feedback regarding cancer outcomes and discriminative skills, the effect of volume may be to simply encourage more positive calls." • "Direct feedback of performance characteristics coupled with training…may be more helpful than experience without feedback." **Author Reported Limitations** • "…the surveyed radiologists were not a random sample of all radiologists in the United States but only a sample participating in the national Breast Cancer Surveillance	litigation concerns) • Data on reading volume are self-reported • Cancer detection rates or PPVs are not reported; therefore, data on sensitivity has been extracted • Final model: only sensitivity and specificity are reported

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				1001-2000: 1.57 (1.09, 2.27) >2000: 1.80 (1.27, 2.54) P=0.004 % of mammograms interpreted in the past year that were screening mammograms: <50: 1.00 (ref.) 51-75: 1.30 (0.79, 2.16) 76-100: 1.30 (0.79, 2.13) P=0.56) **•Sensitivity (95% CI); results from the final model adjusted for radiologist's characteristics** No. of mammograms interpreted in the past year: ≤1000: 1.00 (ref.) 1001-2000: 1.68 (1.18, 2.39) >2000: 1.89 (1.36, 2.63) P=0.001 **•Specificity (95% CI); results from the final model adjusted for radiologist's characteristics** No. of mammograms interpreted in the past year:	Consortium in three distinct locations." • "…reported mammographic volume may have been estimated inaccurately by the radiologists when they responded to the survey." • "…the radiologist is reporting all mammograms in the survey, not just screening mammograms." • "…the volume is reported in discrete categories used in the survey, rather than the actual numbers."	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				≤1000: 1.00 (ref.) 1001-2000: 0.65 (0.52, 0.83) >2000: 0.76 (0.60, 0.96) P=0.002		
Buist, 2011 [USA]	• N/A	• Program: six mammography registries contributing to BCSC • Study period: 2002-2006 • Target age: 40-79 years • Screening frequency: "each year" • Sample size (# of screens) = 783,965; (# of women) = 476,079; (# of radiologists) = 120	• Factor of Study: reading volume • Other potential influencing factors: **Reading Approach:** mammograms were included in the analysis only if the radiologist was the primary reader. **Readers' Training:** radiologists; 92% had no fellowship training in breast imaging **Prior Mammograms:** available for 86% of examinations **Technology:** not reported	• **No. of women recalled per cancer detected** Annual total volume <480: 19.3 480-999: 27.0 1000-1499: 25.9 1500-1999: 20.8 2000-2999: 20.5 3000-4999: 20.2 ≥5000: 23.5 Annual screening volume <480: 23.2 480-999: 27.5 1000-1499: 24.8 1500-1999: 19.2 2000-2999: 20.6 ≥3000: 22.2 Annual diagnostic volume: <100: 17.2 100-199: 17.3 200-299: 20.9 300-499: 22.8 500-999: 25.3 ≥1000: 23.7 Screening focus (%)	**Author Reported Conclusions** • "We found that higher interpretive volume was associated with clinically and statistically important lower rates of false-positive results and numbers of women recalled per cancer detected - without a corresponding decrease in sensitivity or CDR. We also observed lower CDRs in radiologists with low diagnostic volumes." • "Performance across radiologists within volume levels had wide, unexplained variability, reinforcing the ideas that the volume-performance relationship is complex and several factors may influence it."	• Although total reading volume included diagnostic mammograms, performance data were collected only for screening mammograms interpreted within BCSC facilities • Characteristics of screened women did not differ by radiologist total volume (see table 2 of the publication).

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				<75: 23.8 75-79: 27.0 80-84: 23.6 85-89: 22.6 ≥90: 14.5 • **False-positive rate (95% CI) (%)** Annual total volume <480: 7.7 (4.8, 12.1) 480-999: 11.0 (9.0, 13.4) 1000-1499: 11.2 (9.4, 13.3) 1500-1999: 8.3 (7.0, 9.9) 2000-2999: 8.3 (7.1, 9.6) 3000-4999: 8.4 (7.2, 9.7) ≥5000: 9.5 (7.0, 12.7) Annual screening volume <480: 9.9 (6.4, 15.0) 480-999: 11.2 (9.6, 13.0) 1000-1499: 10.6 (9.1, 12.3) 1500-1999: 7.7 (6.4, 9.2) 2000-2999: 8.3 (7.4, 9.4) ≥3000: 9.1 (7.2, 11.4) Annual diagnostic volume: <100: 6.7 (5.4, 8.4) 100-199: 6.8 (5.6, 8.3) 200-299: 8.4 (7.3, 9.8) 300-499: 9.5 (8.0, 11.1) 500-999: 10.5 (9.0, 12.2) ≥1000: 9.8 (7.4, 12.8) Screening focus (%) <75: 9.9 (7.4, 13.2)	• "Radiologists with a greater screening focus had significantly lower sensitivities and CDRs and significantly lower false-positive rates." • "There is no single "best" performance metric that can be used to help set policy. Our simulation results demonstrate that changing MQSA volume requirements or adding minimum numbers of screening and diagnostic examinations could result in modest improvements in some screening outcomes at a cost to others." **Author Reported Limitations** • "Statistical variability issues complicate measuring volume-performance outcomes. Cancer is rare in screening settings…. Because false-negative cases are rare (one per 1000 mammograms) and some are visible only in retrospect…it	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				75-79: 11.6 (10.0, 13.4) 80-84: 9.7 (8.4, 11.2) 85-89: 9.1 (7.8, 10.7) ≥90: 5.6 (4.4, 7.0) • **Cancer Detection Rate per 1000 (95% CI)** Annual total volume <480: 3.4 (2.4, 4.7) 480-999: 4.3 (3.6, 5.1) 1000-1499: 4.6 (4.2, 5.1) 1500-1999: 4.2 (3.5, 5.0) 2000-2999: 4.4 (3.9, 4.9) 3000-4999: 4.7 (3.9, 5.5) ≥5000: 3.6 (3.1, 4.2) Annual screening volume <480: 4.2 (3.2, 5.5) 480-999: 4.3 (3.7, 4.9) 1000-1499: 4.9 (4.4, 5.5) 1500-1999: 4.1 (3.5, 4.7) 2000-2999: 4.4 (3.9, 5.0) ≥3000: 4.0 (3.4, 4.6) Annual diagnostic volume: <100: 3.3 (2.6, 4.1) 100-199: 3.7 (3.2, 4.2) 200-299: 4.3 (3.7, 4.9) 300-499: 4.6 (4.0, 5.3) 500-999: 4.6 (3.9, 5.3) ≥1000: 4.1 (3.6, 4.7) Screening focus (%) <75: 4.5 (3.8, 5.4)	could take many years for a low-volume reader to miss a finding that an expert might identify. This is a smaller problem for high-volume readers." • "We could not explore the influence of feedback."	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				75-79: 5.1 (4.5, 5.8) 80-84: 4.2 (3.8, 4.7) 85-89: 4.2 (3.7, 4.8) ≥90: 3.4 (2.7, 4.2)		
Coldman, 2006 **[Canada]**	- N/A	- Program: data from Canadian provincial screening programs that agreed to participate (Alberta, British Columbia, Manitoba, Newfoundland, Nova Scotia, Ontario, Quebec) - Study period: 1998-2000 - Target age: overall from 40 to 79 years [based on patients' age ranges reported]. All programs target women 50-69 years and some provide screening to women 40-49 or 70-79 years. - Screening frequency:	- Factor of Study: reading volume - Other potential influencing factors: **Reading Approach:** not reported **Readers' Training:** radiologists **Prior Mammograms:** not reported **Technology:** not reported	- **Recall Rate (%)** [termed "abnormal interpretation rate" in the publication] 480–699: 1.00 (ref.) 700–999: 1.03 (0.90, 1.13) 1000–1499: 0.99 (0.90, 1.16) 1500–1999: 1.20 (1.01, 1.40) 2000–2999: 0.97 (0.75, 1.19) 3000–4999: 0.97 (0.83, 1.09) ≥5000: 0.91 (0.73, 1.15) - **Cancer Detection Rate per 1000** 480–699: 1.00 (ref.) 700–999: 1.07 (0.94, 1.22) 1000–1499: 1.02 (0.90, 1.14) 1500–1999: 1.11 (0.95, 1.32) 2000–2999: 1.20 (1.01, 1.38) 3000–4999: 1.13 (0.99, 1.30) ≥5000: 0.99 (0.82, 1.15) - **PPV (%)** 480–699: 1.00 (ref.) 700–999: 1.05 (0.89, 1.23) 1000–1499: 1.07 (0.91, 1.26) 1500–1999: 1.13 (0.90, 1.40) 2000–2999: 1.34 (1.07, 1.65)	**Author Reported Conclusions** - "Radiologist reading volume had no consistent effect on either the cancer detection rate or the abnormal interpretation rate…" - "Our study indicate that inter-radiologist variation was one of the strongest influences on the abnormal interpretation rate." - "…there was a consistent pattern of increasing PPV with higher volumes…" - "…the PPV increased with the volume of screening examinations interpreted up to about 2000 annual screening examinations but then stabilized." - "There was a trend, as identified by the PPV, for radiologists with higher reading volumes	- Radiologists who interpreted fewer than 480 screens per year over the study period were not included in these analyses. - Variables included in each analysis: age, screening sequence, province, average radiologist's volume, inter-radiologist effect.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		- Sample size (# of screens) = 1,543,331; (# of radiologists) = 584 - Mean age (years): not reported [patient's distribution by 10-year age intervals and province is reported in table 2 of the publication]		3000–4999: 1.36 (1.07, 1.61) ≥5000: 1.37 (1.06, 1.84) "Our study indicate that inter-radiologist variation was one of the strongest influences on the abnormal interpretation rate."	to be better able to select for further investigation women who were likely to have breast cancer. This is important because in any clinical situation one wishes to minimize harms, both to the patient in terms of anxiety and to the health system in terms of cost, while providing the maximum benefit. The requirement by some Canadian screening programs of minimum annual volumes that are higher than the 480 mammograms specified by the Canadian Association of Radiologists is supported by the results of this analysis." **Author Reported Limitations** - "Data were not available on other variables that may affect radiologist performance, such as recent education and	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					practice characteristics." - "Screening volumes were estimated by using only data from screening programs, but in some provinces, radiologists may be involved in screening outside a program. Any bias in volumes would most likely cause effects to be underestimated, although true confounding cannot be ruled out."	
Cornford, 2011[11] [UK]	- N/A	- Program: East Midlands Breast Screening Programme - Study period: April 2005-March 2008 - Target age: not reported - Screening frequency: not reported	- Factor of Study: reading volume - Other potential influencing factors: **Reading Approach**: double-read with either consensus or arbitration for discordant cases; the second readers were not blinded to the	Median (range) - **Recall Rate (%)** <15,000/3 years: 6.9 (2.6, 10.4) 15,000 to <20,000/3 years: 6.4 (3.9, 8.7) 20,000 to <25,000: 5.1 (2.3, 7.4) ≥25,000: 3.1 (1.6, 6.9) P=0.053 - **Cancer Detection Rate per 1000**	**Author Reported Conclusions** - "The low volume readers in the study had the highest median recall rate; however, this did not differ significantly from the other reading groups combined, neither was there any significant difference in terms of cancer detection."	- The units of recall rates are unclear: reported by the study authors as rates per 1000 but the values suggest these are rates per 100. - It appears that the analyses were not adjusted for patient's or radiologist's characteristics

[11] Similar results and conclusions are reported by Cornford et al. 2009 (conference abstract): Cornford E, Reed J, Murphy A, Evans A, Bennett R. Optimal mammography reading volumes: evidence from real life. Breast Cancer Res. 2009; 11(Suppl 2): O2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4284827/

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Sample size (# of screens) not reported; (# of film readers) = 37 (N=16 radiographers; N=21 radiologists)	results of the first reader **Readers' Training**: Radiographers with median film reading experience of 5.5 years (range 2-12 years) and median volume of film read during 3-year period of 13,163 (range 9864-19329) Radiologists with median film reading experience of 10 years (range 3-19 years) and median volume of film read during 3-year period of 22,538 (range 4,423-38,632) • All rates were calculated using first-reader data **Technology**: unclear (the terms "film readers" and "film reading" are used throughout the text)	<15,000/3 years: 7.5 (5.8, 10.5) 15,000 to <20,000/3 years: 7.9 (7.1, 9.7) 20,000 to <25,000: 8.3 (7.3, 9.3) ≥25,000: 6.9 (5.4, 8.6) P=0.013 • "The median cancer-detection rate in the high-volume group (≥25,000 mammograms/3 years) was significantly lower than the other groups combined (p= 0.004, 6.9 per 1000 women screened versus 7.9 per 1000 women screened)." **Small Cancer Detection Rate per 1000** <15,000/3 years: 3.9 (2.1, 5.8) 15,000 to <20,000/3 years: 4.6 (3.5, 5.0) 20,000 to <25,000: 4.4 (3.8, 5.2) ≥25,000: 3.8 (2.9, 4.4) • **PPV (%)** <15,000/3 years: 14.2 (5.9, 24.7) 15,000 to <20,000/3 years: 12.5 (9.1, 25.0)	• "…this preliminary study from the East Midlands region has not provided any evidence for reducing the threshold volume of 5000 cases/year. Further the results suggest that there may be reading volumes above which overall cancer detection rates decline. However, at these higher volumes small cancer-detection rates remained comparable with lower-volume readers, and recall rates were lower." **Author Reported Limitations** • The results are likely to be affected by occupational group, with the majority of the lowest volume group being advanced practitioners." • Small sample size • "The present study was not large enough to examine the relationship between	(likely due to small sample size)

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				20,000 to <25,000: 16.4 (11.7, 35.5) ≥25,000: 19.1 (10.7, 36.0)	occupational group and either volume or experience"	
Duncan, 2011 [UK]	• N/A	• Program: the study included performance data of film readers in Scotland • Study period: 2006-2009 • Target age: not reported • Screening frequency: not reported • Sample size (# of readers) = 37 • Mean age (years): not reported	• Factor of Study: reading volume • Other potential influencing factors: **Reading Approach:** "using a paper record and the second reader was not blinded to the first reader's decision, so reading behavior and performance as a second reader is likely to have been affected by the results of the first reader. Therefore, the analysis was based on performance as a first reader, but related to the total number of cases read…"	• **Recall Rate** Mean (SD) *Low (>15,000/3 years)*: 0.05 (0.01) *Medium (15,000-25,000/3years):* *0.05 (0.01)* *High (>25,000/3-years)*: 0.06 (0.01) Median *Low (>15,000/3 years)*: 0.05 *Medium (15,000-25,000/3years):* 0.05 *High (>25,000/3-years)*: 0.06 **P value (ANOVA)=0.37** • **Sensitivity; mean (SD)** Mean (SD) *Low (>15,000/3 years)*: 0.94 (0.03) *Medium (15,000-25,000/3years):* 0.93 (0.06)	**Author Reported Conclusions** • "…there was no evidence of poorer performance above a threshold of 25,000 cases read over a 3-year period…" • "…the present results do not support the suggestion that reading performance drops off with a 3-year case volume of greater than 25,000. However, similar to the study by Cornford et al., the number of readers involved is small." **Author Reported Limitations** • "at the time of writing, the study by Cornford et al. was unpublished, details of their analysis methods are not known and so this study is not directly comparable." • "the number of readers is relatively small…"	• The aim was to investigate whether the finding of Cornford et al. obtained in East Midlands, could be replicated in Scotland. • Readers were divided into high, medium or low experience using thresholds similar to those used by Cornford et al. to investigate whether the performance dropped above 25,000 screens/3 years. • Although readers of different qualification participated in this study, the differences were not accounted for in the analysis.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Readers' Training: radiologists, breast clinicians and radiographers **Prior Mammograms:** not reported **Technology:** film •	*High (>25,000/3-years):* 0.93 (0.05) Median *Low (>15,000/3 years*): 0.94 *Medium (15,000-25,000/3years)*: 0.93 *High (>25,000/3-years)*: 0.94 **P value (ANOVA)=0.28**		• The analysis does not appear to be adjusted for patients' characteristics or for other reader's characteristics (possibly due to small sample size). It is stated that reader's levels of experience (number of years of film reading) were evenly distributed among the low, medium and high-volume reading groups.
Elmore, 2009 [USA]	• N/A	• Program: seven Breast Cancer Surveillance Consortium (BCSC) sites • Study period: January 1, 1998 to December 31, 2005 • Target age: ≥40 years • Screening frequency: not reported	• Factor of Study: reading volume • Other potential influencing factors: **Reading Approach:** not reported **Readers' Training:** radiologists (8% fellowship trained) **Technology:** not reported	Adjusted ORs (95% CI) • **Recall Rate** Self-reported average no. of mammograms interpreted per year over the past 5 years ≤1000: 1.00 (ref.): 1001-2000: 1.19 (0.92, 1.55) >2000: 1.03 (0.79, 1.33) P=0.170 % of images from all examinations that were screening mammograms (annual average over past 5 years)	**Author Reported Conclusions** • "We found no association between self-reported annual volume of mammograms interpreted and interpretive performance." **Author Reported Limitations** • "…low numbers of examinations in women with cancer… added to	• Self-reported reading volume • Adjustment for patients' characteristics (BCSC registry, age, breast density, time since last mammographic examination), radiologists' random effect and radiologists' characteristics (gender, affiliation, experience)

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Sample size (# of screens) = 1,036,155; (# of women) = 531,705; (# of radiologists) = 205 • 257 of 364 eligible radiologists responded to a self-administered mail survey (71% response rate); survey results were linked to BCSC data on screening mammograms interpreted by these radiologists. Twenty-six radiologists with incomplete BCSC data were excluded. •		<83%: 1.00 (ref.) ≥83%: 0.98 (0.84, 1.15) P=0.812 • **False positive rate** Self-reported average no. of mammograms interpreted per year over the past 5 years ≤1000: 1.00 (ref.): 1001-2000: 1.19 (0.90, 1.56) >2000: 1.03 (0.79, 1.35) P=0.210 % of images from all examinations that were screening mammograms (annual average over past 5 years) <83%: 1.00 (ref.) ≥83%: 0.98 (0.84, 1.15) P=0.833 • **PPV1** Self-reported average no. of mammograms interpreted per year over the past 5 years ≤1000: 1.00 (ref.): 1001-2000: 0.93 (0.68, 1.27) >2000: 0.94 (0.70, 1.28) P=0.910	the variability we found in sensitivity • "…small number of fellowship-trained radiologists (n = 16) and the lack of data on the use of digital mammography." • "…30% of the study radiologists interpreted mammograms at institutions outside of the BCSC; thus, their self-reported data on annual volume could not be verified." • "…many of the radiologists worked part time, and this factor made interpretation of the percentage of time spent in breast imaging challenging.	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				% of images from all examinations that were screening mammograms (annual average over past 5 years) <83%: 1.00 (ref.) ≥83%: 0.96 (0.82, 1.14) P=0.667 • **Sensitivity** Self-reported average no. of mammograms interpreted per year over the past 5 years ≤1000: 1.00 (ref.): 1001-2000: 1.12 (0.64, 1.97) >2000: 0.87 (0.51, 1.48) P=0.255 % of images from all examinations that were screening mammograms (annual average over past 5 years) <83%: 1.00 (ref.) ≥83%: 0.96 (0.73, 1.26) P=0.766		
Smith-Bindman, 2005 [USA]	• N/A	• Program: data from four mammography registries participating in the Breast Cancer	• Factor of Study: • Other potential influencing factors: **Reading volume:**	• **Specificity (OR, 95% CI)** Average annual volume of mammograms 481-750: 1.0 (ref.)	**Author Reported Conclusions** • "In general, the most experienced physicians had the lowest false-positive rates.	• Adjustment for patient's characteristics and physician's characteristics

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		Surveillance Consortium (BCSC) • Study period: January 1, 1995 to December 31, 2000 • Target age: not reported • Screening frequency: not reported • Sample size (# of screens) = 1,220,046' (# of physicians) = 209 • Mean age (years): not reported	Only physicians who read >=480 mammograms per year were included. **Reading Approach:** **Readers' Training:** "physicians" **Prior Mammograms: not reported** • **Technology: not reported**	751-1000: 1.14 (0.93 to 1.41). P=0.216 1001-1500: 1.05 (0.85 to 1.30). P=0.657 1501-2500: 1.16 (0.97 to 1.39). P=0.092 2501-4000: 1.30 (1.06 to 1.59). P=0.011 >4000: 1.03 (0.85 to 1.25). P=0.789 Ratio of screening to diagnostic mammograms <5: 1.0 (ref.) >5: 1.59 (1.37 to 1.82). P<0.001 • **Sensitivity (OR, 95% CI)** Average annual volume of mammograms 481-750: 1.0 (ref.) 751-1000: 1.17 (0.87 to 1.56). P=0.292 1001-1500: 1.07 (0.80 to 1.44). P=0.643 1501-2500: 0.91 (0.72 to 1.15). P=0.449 2501-4000: 0.83 (0.63 to 1.10). P=0.197 >4000: 0.96 (0.74 to 1.23). P=0.719 Ratio of screening to diagnostic mammograms	Physicians who had been practicing the longest, who interpreted 2500 – 4000 mammograms annually, and who emphasized screening, as opposed to diagnostic, mammography had lower false-positive rates than their less experienced counterparts. For physicians who had practiced the longest and who had a high focus on screening mammography, overall accuracy was improved as well, meaning that they had higher specificity without an equal loss in sensitivity." **Author Reported Limitations** • "…we do not know whether greater experience, higher annual volume, and a greater focus on screening mammography improve interpretations or whether the better	• This article also reports on the effect of physician's age and time since receipt of medical degree

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				<5: 1.0 (ref.) >5: 0.82 (0.69 to 0.98). P=0.026 **Accuracy (OR, 95% CI)** Average annual volume of mammograms 481-750: 1.0 (ref.) 751-1000: 1.33 (0.97 to 1.83). P=0.080 1001-1500: 1.13 (0.87 to 1.46). P=0.373 1501-2500: 1.06 (0.86 to 1.32). P=0.571 2501-4000: 1.08 (0.82 to 1.42). P=0.586 >4000: 0.98 (0.77 to 1.25). P=0.878 Ratio of screening to diagnostic mammograms <5: 1.0 (ref.) >5: 1.29 (1.08 to 1.55). P=0.005	physicians simply choose to interpret more examinations." • Sample size "was not large enough to look separately at ductal carcinoma in situ and invasive cancer".	
Theberge, 2005 [Canada]	• N/A	• Program: Quebec Breast Cancer Screening Program • Study period: May 1998-December 2000	• Factor of Study: • Other potential influencing factors: **Reading Approach:** not reported (inability to account for double reading is	Adjusted rate ratios by annual reading volume (95% CI) • **False-positives** Radiologist's volume <250: 1.00 (ref.) 250-499: 1.02 (0.78, 1.35) 500-749: 0.98 (0.73, 1.31) 750-999: 1.01 (0.75, 1.35)	**Author Reported Conclusions** • "The radiologists' caseload did not seem to influence their ability to detect invasive or in situ breast cancer." • "By contrast, cancer detection was	• The data were analysed using a case-control approach: 1) cancer detection rates were analyzed by comparing 1709 women with

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Target age: 50-69 years • Screening frequency: not reported • Sample size (# of women) = 307,314; (# of radiologists) = 275; (# of facilities) = 68 • Mean age (years): Women with screen-detected breast cancer 59.5 (5.9) Women without screen-detected cancers abnormal mammogram 57.8 (5.8) normal mammogram 58.3 (5.7)	listed as a limitation of the study) **Readers' Training:** radiologists **Prior Mammograms:** not reported **Technology:** not reported	1000-1249: 0.89 (0.63, 1.26) 1250-1499: 0.93 (0.65, 1.33) ≥1500: 0.53 (0.35, 0.79) P value for trend: 0.001 Facility's volume <2000: 1.00 (ref.) 2000-2999: 0.91 (0.73, 1.15) 3000-3999: 0.95 (0.73, 1.24) ≥4000: 1.20 (0.94, 1.51) P value for trend: 0.09 Radiologist's annual volume in facilities with annual volume <3000 <499: 1.00 (ref.) 500-999: 1.06 (0.85, 1.31) ≥1000: 0.85 (0.65, 1.12) Radiologist's annual volume in facilities with annual volume ≥3000 <499: 1.29 (1.03, 1.61) 500-999: 1.12 (0.92, 1.36) ≥1000: 0.92 (0.68, 1.24) • **Cancer Detection Rate** Radiologist's volume <250: 1.00 (ref.) 250-499: 1.00 (0.73, 1.38) 500-749: 0.93 (0.68, 1.28) 750-999: 1.05 (0.76, 1.45) 1000-1249: 1.05 (0.76, 1.46) 1250-1499: 1.06 (0.75, 1.48)	associated with the number of screenings performed in the facility." • "The false-positive rate ratio decreased significantly …with increasing screening volume of radiologists." • "By contrast, the screening volume of facilities was not associated with false-positive readings…" • "Radiologists who worked in facilities performing a greater number of screenings per year had higher detection rates than those who worked in facilities performing fewer, and this was true for all radiologists working in high-volume facilities, irrespective of their individual screening volume. In contrast, the false-positive rates decreased with increasing radiologist caseload, and this trend was clearer among those who worked in larger facilities.	screen-detected breast cancer and a 10% random sample (n=30,560) of women without screen-detected cancer. 2) False-positive rates were analyzed by comparing 3,159 women with false-positive readings and 27,401 other women in the random sample. The odds ratios from these analyses approximated the cancer detection rate ratios and the false-positive rate ratios. • Analyses were adjusted for patient's age, BMI, previous mammography, biopsy or implant, abnormal recall rate of radiologist's colleagues, radiologist's year of certification and number of

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				≥1500: 0.97 (0.67, 1.41) P value for trend: 0.71 Facility's volume <2000: 1.00 (ref.) 2000-2999: 0.96 (0.82, 1.11) 3000-3999: 1.10 (0.94, 1.28) ≥4000: 1.28 (1.07, 1.52) P value for trend: 0.004 Radiologist's annual volume in facilities with annual volume <3000 <499: 1.00 (ref.) 500-999: 0.99 (0.84, 1.17) ≥1000: 1.08 (0.92, 1.28) Radiologist's annual volume in facilities with annual volume ≥3000 <499: 1.20 (0.98, 1.48) 500-999: 1.20 (1.02, 1.43) ≥1000: 1.25 (1.03, 1.52)	• "In our analysis, radiologists who read larger numbers of screening mammograms and worked in facilities performing larger numbers of screenings tended to have higher breast cancer detection rates while maintaining lower false-positive rates than radiologists who performed fewer readings and worked in facilities performing fewer screenings." • "In conclusion, radiologists' and facilities' screening volumes appear to have independent and complimentary influences on performance as measured by rates of breast cancer detection and false-positive readings." • "…the overall performance of screening mammography seems to be maximized when screenings are performed in larger	practices. The analyses of radiologist's screening volume were adjusted for facilities' volumes and vice versa.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					centres and when, in these centres, mammograms are read by radiologists who interpret a large volume of films." **Author Reported Limitations** • "…there were few radiologists or facilities with large numbers of screenings…" • "…our study covered only the first 2 years of the PQDCS, which might not be representative of current functioning." • "only screening-mammography volumes of radiologists and facilities were available…" • "…other important determinants of performance, such as double reading, participation in teaching or research, daily quality-control procedures within facilities and specific training of the radiologists in the interpretation of	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					screening mammograms, could not be taken into account.”	
Theberge, 2014 [Canada]	• N/A	• Program: the Quebec Breast Cancer Screening Program • Study period: 2000-2006 • Target age: 50-69 years • Screening frequency: biannual • Sample size (# of screens) = 1,315,327; (# of women) = 644,498); (# of radiologists) = 340 • Mean age (years): Women with screen-detected breast cancer 58.9 (5.6) Women without screen-detected cancers	• Factor of Study: reading volume • Other potential influencing factors: **Reading Approach:** not reported **Readers’ Training:** radiologists **Prior Mammograms:** **Technology** •	Adjusted ratios (95% CI) • **False positives** Total annual volume <500: 1.11 (1.00 to 1.22) 500-999: 1.00 (ref.) 1000-1499: 0.91 (0.84 to 0.98) 1500-1999: 0.85 (0.77 to 0.95) 2000-2999: 0.82 (0.73 to 0.93) 3000-3999: 0.78 (0.67 to 0.90) ≥4000: 0.76 (0.65 to 0.89) P for trend: 0.001 Screening volume <500: 1.13 (1.05–1.22) 50-999; 1.00 (ref.) 1000-1499: 0.94 (0.87 to 1.01) 1500-1999: 0.90 (0.80 to 1.01) 2000-2499: 0.80 (0.70 to 0.91) ≥2500: 0.85 (0.73 to 0.99) P for trend: 0.006 Diagnostic volume <500: 1.06 (0.97 to 1.16) 500-999: 1.00 (ref.) 1000-1499: 0.90 (0.82 to 0.99) ≥1500: 0.81 (0.72 to 0.91) P trend: <0.001	**Author Reported Conclusions** • “In this Canadian organized mammography screening program, an increase in annual interpretive volume was associated with little or no change in sensitivity but with reductions in false-positive rates. Thus, screening accuracy (sensitivity/false-positive rate) increased with increasing volume.” • “Although our data suggest that accuracy increases throughout the range of annual volumes observed in this study (up to approximately 6000 mammograms annually), the change in accuracy associated with increasing volume seems to be greater up	• All models were adjusted for characteristics of patients (age, BMI, breast density, family history of breast cancer, postmenopausal status, parity, HRT, clinical breast examination in the past year, previous breast aspiration or biopsy, screening history), radiologist’s characteristics (sex, year of graduation, medical school attended) and facilities characteristics (facility type and volume)

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		abnormal mammogram 56.9 (5.6) normal mammogram 58.0 (5.5) •		"Smoothed plots showed a greater reduction in false-positive rates at the lower volume (for all volume types), with the curve stabilizing at higher volume." • **Sensitivity** Total annual volume <500: 0.98 (0.95 to 1.02) 500-999: 1.00 (ref.) 1000-1499: 0.99 (0.95 to 1.02) 1500-1999: 1.00 (0.96 to 1.03) 2000-2999: 1.01 (0.98 to 1.04) 3000-3999: 0.99 (0.96 to 1.02) ≥4000: 1.00 (0.96 to 1.04) P for trend: 0.68 Screening volume <500: 1.00 (0.97 to 1.02) 50-999; 1.00 (ref.) 1000-1499: 1.00 (0.98 to 1.03) 1500-1999: 1.03 (1.00 to 1.06) 2000-2499: 1.02 (0.99 to 1.05) ≥2500: 1.00 (0.96 to 1.04) P for trend: 0.87 Diagnostic volume <500: 1.01 (0.98 to 1.03) 500-999: 1.00 (ref.) 1000-1499: 1.03 (1.00 to 1.06) ≥1500: 1.01 (0.97 to 1.05) P for trend: 0.80	to 3000 mammograms per year. Gains in accuracy beyond 3000 mammograms annually are minimal." • "In conclusion, this study suggests that the minimal volume requirement of 500 mammograms annually adopted in North America is justified. Radiologist accuracy may be compromised when interpretive volume consistently falls short of this minimum requirement. Raising the interpretive volume of radiologists may help to minimize false-positive screens without sacrificing sensitivity. Our results demonstrate that potential gains in accuracy with increases in volume may be greater up to an annual interpretive volume of approximately 3000 mammograms." **Author Reported Limitations**	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				"Spline analyses showed little change in sensitivity with increases in total volume… Variations in sensitivity with total volume were not statistically significantly different from a straight horizontal line." • **Accuracy (=sensitivity/false-positive rate)** <u>Total annual volume</u> <500: 0.89 (0.80 to 0.99) 500-999: 1.00 (ref.) 1000-1499: 1.08 (1.00 to 1.18) 1500-1999: 1.17 (1.04 to 1.30) 2000-2999: 1.23 (1.09 to 1.38) 3000-3999: 1.27 (1.10 to 1.47) ≥4000: 1.32 (1.13 to 1.54) P for trend: 0.0005 <u>Screening volume</u> <500: 0.88 (0.81 to 0.96) 50-999; 1.00 (ref.) 1000-1499: 1.07 (1.00 to 1.15) 1500-1999: 1.15 (1.02 to 1.28) 2000-2499: 1.27 (1.12 to 1.44) ≥2500: 1.18 (1.02 to 1.37) P for trend: 0.003 <u>Diagnostic volume</u> <500: 0.95 (0.87 to 1.04)	• We could not adjust for fellowship training of radiologist, which seems to be a possible confounding factor… However, in Quebec, only a small proportion of radiologist's complete fellowship training in mammography, thus reducing the extent of possible confounding." • Some misclassification of mammograms as screening or diagnostic was possible.	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				500-999: 1.00 (ref.) 1000-1499: 1.14 (1.03 to 1.25) ≥1500: 1.25 (1.11 to 1.39) P for trend: <0.001 "Smoothed plots showed an increasing trend in accuracy with increasing volume for all volume types…" "Increases in total volume are associated with a greater increase in accuracy up to approximately 3000 mammograms per year."		
			Double Reading			
Roman, 2012 **[Spain]**	- N/A	- Program: population-based screening program in Spain - Study period: March 1990 to December 2006 - Target age: 50-69 years - Screening frequency: every 2 years - Sample size (# of screens) = 4,739,498;	- Factor of Study: double vs. single reading - Other potential influencing factors: **Reading Approach**: 84.8% of double readings involved consensus or arbitration; 15.2% were double readings without consensus	Adjusted OR (95% CI) - **False-positive risk (all procedures)** Single reading: 1.00 (ref.) Double reading 2.06 (2.00, 2.13) - **False-positive risk (invasive procedures)** Single reading: 1.00 (ref.) Double reading 4.44 (4.08, 4.84) - **Cancer Detection** Single reading: 1.00 (ref.)	**Author Reported Conclusions** - "…we found that double reading was associated with a higher recall rate… and a higher cancer detection rate… than single reading." **Author Reported Limitations** - **"**The information on women's personal variables was not always available or complete in all the radiology units."	- Cited as Almazan et al. 2012 in the list of original publications - Adjustment for women's screening number, radiology unit (random effect), screening period (calendar years) and age - Unclear whether the second reader was aware of the first reader's interpretation

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		(# of women) = 1,565,364 - Mean age (years):	**Readers' Training:** radiologists **Prior Mammograms:** not reported **Technology:** not reported	Double reading 1.08 (1.04, 1.12)		
Bennett, 2012 [UK]	- Radiographers as mammogram readers	--	--	--	**--**	--
Caumo, 2011a[12] **[Italy]**	- N/A	- Program: a mammography screening program of an Italian Local Health Unit - Study period: March 2007- October 2008 - Target age: 50-69 years - Screening frequency: not reported - Sample size (# of women) = 23,639	- Factor of Study: delayed second reading procedure as an adjunct to real-time reading with immediate assessment. - Other potential influencing factors: **Reading Approach:** - "first reading was…associated with immediate assessment and second reading followed in a separate session."	- **Recall rate at first reading (% screened)** 13.0% - **Recall rate at second reading only (% screened)** 2.7% - **Recall rate after both readings (% screened)** 15.7% - **Incremental Recall Rate at second reading (% change relative to first reading)** +21.1% - **PPV of recall at first reading (%):** 5.4	**Author Reported Conclusions** - "In conclusion, our study findings confirm the usefulness of second reading of screening mammograms and the limitations of single reading, even if read in real time with immediate assessment. As a consequence of these findings, in this scenario, the real-time reading plus immediate assessment policy has been abandoned in	- Unilateral recall - Unblinded second reading - Variables considered: patient's age, breast density

[12] Caumo, F., Brunelli, S., Zorzi, M., Baglio, I., Ciatto, S., & Montemezzi, S. (2011). Benefits of double reading of screening mammograms: retrospective study on a consecutive series. Radiologia Medica, 116(4), 575-583

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Mean age (years): not reported; age 50-59 years: 46.6% age 60-69 years: 53.4%	• "Double reading was informed and not independent, the second reader being aware of the first reader's report." **Readers' Training:** radiologists; two expert readers with >5 years of experience and one with one year of experience. **Prior Mammograms:** **Technology:** full-field digital mammography (see Caumo et al. 2011b)	• **PPV of recall at second reading only (%): 3.3** • **Cancer detection rate at first reading (per 1000 screened)** 7.06 • **Cancer detection rate at second reading only (per 1000 screened)** 0.93 • **Cancer detection rate after both readings (per 1000 screened)** 7.99 • **Incremental Cancer Detection Rate per 1000 (% change relative to first reading)** +13.1%	favour of conventional delayed double reading, as in the rest of Italian screening programmes." **Author Reported Limitations** • None reported •	
Caumo, 2011b[13] [Italy]	• N/A	• Program: Verona and Padua mammography screening programs • Study period: 15 September 2009 – 15 January 2010	• Factor of Study: the effect of arbitration of discordant opinions in double reading • Other potential influencing factors:	Estimates • **Number of assessment procedures spared by arbitration** 216 • **Absolute reduction of recall rate by arbitration of discordant opinions (%)**	**Author Reported Conclusions** • "Arbitration is a cost-effective procedure that could be employed as a first measure to counterbalance excess recall rate observed in a	• It is unclear whether the second reader was aware of the opinion of the first reader. • The study included all recalls to

[13] Caumo, F., Brunelli, S., Tosi, E., Teggi, S., Bovo, C., Bonavina, G., et al. (2011). On the role of arbitration of discordant double readings of screening mammography: experience from two Italian programmes. Radiol Med, 116(1), 84-91.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Target age: see Caumo et al. 2011a • Screening frequency: not reported • Sample size (# of screens) = 7,660 • Mean age (years):	**Reading Approach:** Recalls by one of two readers (discordant) were arbitrated by a third reader with >30 years of experience and >200,00 readings in screening mammography **Readers' Training:** six radiologists with mammography reading experience from 2 to >10 years **Prior Mammograms:** not reported **Technology:** full-field digital mammography •	2.8 • **Relative reduction of recall rate by arbitration of discordant opinions (%)** 40.9 • **Number of cancers missed due to arbitration of discordant cases** 1 • **Absolute reduction of detection rate by arbitration of discordant opinions (per 1000)** 0.13 **Relative reduction of detection rate by arbitration of discordant opinions (%)** 2.0 • "Arbitration cost was 74 euros, whereas 216 spared assessment procedures would have cost 14,558.4-23.346 euros."	double-reading scenario." **Author Reported Limitations** • "…some imprecision of cost estimates might have occurred."	diagnostic assessment during the study period. All recalled cases, irrespective of arbitration results, underwent diagnostic assessment, and the results of these assessments were used as the reference standard for estimation of absolute and relative reductions in recall and cancer detection rates due to arbitration. •
Ciatto, 2005 **[Italy]**	• N/A	• Program: Florence population-based screening program • Study period: January 1998-June 2003	• Factor of Study: double vs. single reading • Other potential influencing factors: **Reading Approach:** double reading, the	• **Recall Rate (%)** <u>First reader</u>: 2.89 <u>Second reader</u>: 3.15 <u>The two readers together</u>: 3.59	**Author Reported Conclusions** • "The contribution of second reading to cancer detection rate in the present retrospective study was rather limited in	• The aim was to evaluate "the impact of second reading subsequent to, and aware of, first reading"

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		- Target age: 50-69 years - Screening frequency: biennial - Sample size (# of screens) = 177,631); (# of radiologists) = 11 - Mean age (years): not reported	second reader is aware of the first reader's report. Referral to assessment is prompted by suspicion by either reader with no consensus or arbitration of discordant cases. **Readers' Training:** radiologists (at least 20,000 screening mammograms read) **Prior Mammograms:** available at request **Technology:** not reported	Additional referral rate introduced by the second reader: +0.70% (24% increase in the first reader's referral rate) Referral rates varied by patient's age - **Cancer Detection** First reader: 670 Second reader: 695 The two readers together: 713 Cancers detected only by one reader: 61 (18 by the first and 43 by the second) Additional cancers detected by the second reader: +0.024% (6.4% increase in cancer detection rate as compared to the first reader) "Detecting 43 additional cancers required 177,631	magnitude, although the cost in terms of extra referrals might still be acceptable." - "…implementing second reading requires a doubling of the number of workload of involved radiologists, which is a crucial aspect, as, unfortunately, radiologists properly trained in reading screening mammograms are currently lacking in Europe". **Author Reported Limitations** - None reported	- "The fact that the expert readers were employed might justify a reduced benefit from double reading, whereas a larger benefit might be expected with less experienced readers."

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				additional readings and 1250 additional referrals." - **PPV for first reader referrals:** 13.04% - **PPV of additional referrals:** 3.44% ["This figure is quite low compared with the positive predictive value… for first reader referrals, but may be easily explained by the fact that cancers missed by the first reader, and thus available for additional detection, are likely to be more difficult to be perceived…"] - **Additional costs:** 2.70 euros per woman screened with double reading; 11,168 euros per additional cancer detected 11,585 euros per cancer detected by single reading		
Gromet, 2008 [USA]	- N/A	- Program: community-based mammography program in Charlotte, NC - Study period: January 1, 2001-December 31, 2005	- Factor of Study: double reading vs. the first reader in a double-reading program - Other potential influencing factors:	- **Recall Rate (%)** First reader [without regard for the second reader's contribution]: 10.2 Final decision [after the second reading and third opinion if required]: 11.9	**Author Reported Conclusions** - "In conclusion, we found that both double reading and CAD are effective methods to increase the sensitivity of screening mammography for	- The study also compares performance before and after CAD implementation -

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Target age: not reported • Screening frequency: not reported • Sample size (# of screens) = 231,221; (# of radiologists) = 9 • Mean age (years):	**Reading Approach:** Batch reading; double reading until 2003; conversion to single reading with CAD during 2003. Cases classified as negative by the first reader and positive by the second reader were resolved by a different subspecialist reader who determined the final reading. **Readers’ Training:** radiologists; first readers were specialized mammographers; second reading was performed by a general radiologist with certification in mammography who did not specialize in the area. Experience: 1-24 years (mean 15 years). The only radiologist with <5	• **Detection Rate per 1000** First reader [without regard for the second reader’s contribution]: 4.12 Final decision [after the second reading and third opinion if required]: 4.46 • **PPV1 (%)** First reader [without regard for the second reader’s contribution]: 4.1 Final decision [after the second reading and third opinion if required]: 3.7 • **Benefit of double reading:** 38 additional cancers detected at a cost of 2,008 additional patients recalled and 140 additional biopsies; PPV decreased to 3.7% and the cancer detection rate increased by 0.34 per 1000; sensitivity increased from 81.4% to 88.0%.	experienced mammogram readers. In our study, the second reader increased sensitivity 6.6%, from 81.4% to 88.0%; the recall rate rose from 10.2% to 11.9%. Single reading enhanced by CAD review yielded a higher sensitivity of 90.4%, with a smaller increase in the recall rate from 10.2% to 10.6%. With manpower and cost constraints limiting the use of double reading in the United States, CAD appears to be an effective alternative that provides similar, and potentially greater, benefits.” **Author Reported Limitations** • The limitation reported (possible effect of improved radiologists’ skills over time on performance) is more relevant to the comparison of periods before and after CAD implementation (see “Technology”)	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			years of experience joined directly after fellowship training. Annual volume from 4,459 to 15,281 readings. **Prior Mammograms:** if available; preference was given to a 3-year old prior examination; additional prior films were available on request **Technology:** screen-film •			
Klompenhouwer, 2015a [the Netherlands]	• N/A	• Program: three screening units in the Southern Netherlands (part pf the Dutch nationwide breast cancer screening program) • Study period: July 2009 to July 2011 • Target age: 50-75 years	• Factor of Study: blinded vs. non-blinded double reading. The reading strategy (blinded and non-blinded) were alternated monthly. • Other potential influencing factors: **Reading Approach:** **D**ouble reading.	• **Recall Rate (%)** Blinded:3.3 Non-blinded: 2.9 P=0.002 • **False-Positive Rate (%)** Blinded: 2.58 Non-blinded: 2.21 P=0.02 • **Detection Rate per 1000** Blinded: 7.4 Non-blinded: 6.5	**Author Reported Conclusions** • "We advocate the use of blinded double reading in order to achieve a better programme sensitivity, at the expense of an increased referral rate and false positive referral rate." **Author Reported Limitations**	•

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Screening frequency: biennial • Sample size (# of screens) = 87,487; (# of radiologists) = 12 • Mean age (years): 59 (95% CI: 59-60)	Women with discrepant readings between the two radiologists, at blinded and non-blinded double reading, were always recalled for further analysis. **Readers' Training:** 12 certified screening radiologists; each evaluated at least 6000 screening mammograms per year **Prior Mammograms:** always available **Technology: F**ull field digital mammography (FFDM)	P=0.139 • **PPV (%)** Blinded:22.1 Non-blinded: 23.1 P=0.507 • **Sensitivity (%)** Blinded:83.1 Non-blinded: 75.5 P=0.003	• "The study design could have benefited from randomization of screened women among the two screening strategies... However, we expect that our quasi-randomised model (screening strategies were altered on a monthly basis) will not result in different outcomes than a true randomisation." • The radiologists "had little experience with FFDM screening at the start of the study. Nevertheless, it is unlikely that our results have been influenced by a learning effect, as the referral rate, cancer detection rate and PPV of referral did not change during the FFDM screening period". • "Our study does not provide information on the cost effectiveness of blinded double reading."	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
Klompenhouwer , 2015b [the Netherlands]	- N/A	- Program: three screening units in the Southern Netherlands (part pf the Dutch nationwide breast cancer screening program) - Study period: 1 July 2009 to 1 July 2011 - Target age: 50-75 - Screening frequency: biennial - Sample size (# of screens) = 84,927 - Mean age (years): 59 (95% CI: 59-60)	- Factor of Study: the effect of arbitration at blinded and non-blinded double reading - Other potential influencing factors: **Reading Approach:** Double reading. The reading strategy (blinded and non-blinded) were alternated monthly. Women with discrepant readings between the two radiologists, at blinded and non-blinded double reading, were always recalled for further analysis. For the purpose of this study, each discordant reading was randomly assigned to a third screening radiologist who retrospectively determined whether he/she would have	- **Recall Rate, % (95% CI)** <u>Blinded double reading</u> Recall of all women with discrepant readings (no arbitration): 3.4 (3.2, 3.5) Recall after arbitration of discrepant readings: 2.2 (2.1, 2.3) P<0.001 <u>Non-blinded double reading</u> Recall of all women with discrepant readings (no arbitration): 2.8 (2.7, 3.0) Recall after arbitration of discrepant readings: 2.3 (2.1, 2.4) P<0.001 - **Detection Rate per 1000 (95% CI)** <u>Blinded double reading</u>	**Author Reported Conclusions** - "Our study showed that discrepant readings occurred significantly more often at blinded double reading. At both blinded and non-blinded double reading, arbitration by a third reader would have resulted in a significantly lower recall rate and significantly higher PPV, without a significant change in the CDR. However, a reduced programme sensitivity would have been obtained after arbitration; this effect was statistically significant at blinded double reading. Expressed in numbers, arbitration at blinded double reading would result in 18.1 less recalls per missed cancers and at non-blinded double reading arbitration would result in 19.3 less recalls per missed cancers." - "Arbitration of discrepant screening	- -

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			recalled the woman. The arbitrator was blinded to the screening outcome. **Readers' Training:** 12 certified screening radiologists; each evaluated at least 6000 screening mammograms per year **Prior Mammograms:** **Technology: F**ull field digital mammography (FFDM) •	Recall of all women with discrepant readings (no arbitration): 7.5 (6.7, 8.3) Recall after arbitration of discrepant readings: 6.8 (6.1, 7.6) P=0.258 Non-blinded double reading Recall of all women with discrepant readings (no arbitration): 6.6 (5.8, 7.4) Recall after arbitration of discrepant readings: 6.3 (5.5, 7.1) P=0.604 • **PPV, % (95% CI)** Blinded double reading Recall of all women with discrepant readings (no arbitration): 22.3 (20.1, 24.4)	mammography assessments is a good tool to improve recall rate and PPV, but is not desirable as it reduces the programme sensitivity at blinded double reading." **Author Reported Limitations** • "We do not provide information on the cost effectiveness of arbitration at blinded versus non-blinded double reading."	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				Recall after arbitration of discrepant readings: 31.2 (28.3, 34.2) P<0.001 Non-blinded double reading Recall of all women with discrepant readings (no arbitration): 23.2 (20.8, 25.6) Recall after arbitration of discrepant readings: 27.5 (24.7, 30.3) P=0.021 **Sensitivity, % (95% CI)** Blinded double reading Recall of all women with discrepant readings (no arbitration): 83.2 (79.5, 87.0) Recall after arbitration of discrepant readings: 76.0 (71.8, 80.3) P=0.013		

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				<u>Non-blinded double reading</u> Recall of all women with discrepant readings (no arbitration): 76.0 (71.5, 80.4) Recall after arbitration of discrepant readings: 72.7 (68.0, 77.3) P=0.308		
Klompenhouwer, 2015c [the Netherlands]	• N/A	• Program: three screening units in the Southern Netherlands (part pf the Dutch nationwide breast cancer screening program) • Study period: 1 July 2009 to 1 July 2011 • Target age: 50-75 • Screening frequency: biennial • Sample size (# of screens) = 84,927	• Factor of Study: the effect of arbitration **<u>on BI-RADS 0 recalls</u>** at blinded and non-blinded double reading • Other potential influencing factors: **Reading Approach:** Double reading. The reading strategy (blinded and non-blinded) were alternated monthly. Women with discrepant readings between the two radiologists, at	• **Recall Rate, % (95% CI)** <u>Blinded double reading</u> Recall of all women with discrepant readings (no arbitration): 3.4 (3.2, 3.5) Recall after arbitration of discrepant <u>BI-RADS 0</u> readings: 2.8 (2.6, 2.9) P<0.001 <u>Non-blinded double reading</u>	**Author Reported Conclusions** • Arbitration of discrepant <u>BI-RADS 0</u> recalls would have significantly lowered recall rate at blinded and non-blinded double reading without a decrease in cancer detection rate and sensitivity. Arbitration would have significantly increased the PPV at blinded double reading. • "…we advise arbitration of discrepant BI-RADS 0 recalls, both at blinded and non-blinded double reading of screening mammograms, to	• "In the Netherlands, women with a screening BI-RADS 0,4 or 5 are recalled and further evaluated at a dedicated hospital breast unit. BI-RADS category 0 represents an abnormality with low suspicion requiring additional work-up…These women have a lower positive predictive value (PPV) of screening

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		- Mean age (years): 59 (95% CI: 59-60)	blinded and non-blinded double reading, were always recalled for further analysis. For the purpose of this study, each discrepant **BI-RADS 0** reading was randomly assigned to a third screening radiologist who retrospectively determined whether he/she would have recalled the woman. The arbitrator was blinded to the screening outcome of the BI-RADS 0 recall. **Readers' Training:** 12 certified screening radiologists with 1-15 years of experience; each evaluated at least 6000 screening	Recall of all women with discrepant readings (no arbitration): 2.8 (2.7, 3.0) Recall after arbitration of discrepant BI-RADS 0 readings: 2.5 (2.4, 2.7) P=0.008 - **Detection Rate per 1000 (95% CI)** Blinded double reading Recall of all women with discrepant readings (no arbitration): 7.5 (6.7, 8.3) Recall after arbitration of discrepant BI-RADS 0 readings: 7.3 (6.5, 8.1) P=0.751 Non-blinded double reading Recall of all women with discrepant readings (no arbitration): 6.6 (5.8, 7.4)	reduce recall rates and improve the PPV of recall at blinded double reading." **Author Reported Limitations** - "…each discrepant BI-RADS 0 reading was re-assessed retrospectively by a third screening radiologist to decide whether or not he/she would have recalled the woman (arbitration). Therefore, the arbiter knew that his or her decision did not have clinical implications for the screening. This lack of clinical implications may have influenced the third reader's decision." - "We provide no information on the cost-effectiveness of arbitration of discrepant BI-RADS 0 recalls at blinded versus nonblinded double reading."	mammography, (14.1%) than BI-RADS 4 (39.1%) and BIRADS 5 (92.9%)… Little is known about the mammographic and tumor characteristics of BI-RADS 0 recalls."

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			mammograms per year **Prior Mammograms:** **Technology: F**ull field digital mammography (FFDM) •	Recall after arbitration of discrepant BI-RADS 0 readings: 6.5 (5.7, 7.2) P=0.832 • **PPV, % (95% CI)** Blinded double reading Recall of all women with discrepant readings (no arbitration): 22.3 (20.1, 24.4) Recall after arbitration of discrepant BI-RADS 0 readings: 26.3 (23.8, 28.3) P=0.015 Non-blinded double reading Recall of all women with discrepant readings (no arbitration): 23.2 (20.8, 25.6) Recall after arbitration of discrepant BI-RADS 0 readings: 25.4 (22.8, 28.0) P=0.213		

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				Sensitivity, % (95% CI) Blinded double reading Recall of all women with discrepant readings (no arbitration): 83.2 (79.5, 87.0) Recall after arbitration of discrepant BI-RADS 0 readings: 81.2 (77.3, 85.1) P=0.453 Non-blinded double reading Recall of all women with discrepant readings (no arbitration): 76.0 (71.5, 80.4) Recall after arbitration of discrepant BI-RADS 0 readings: 74.6 (70.1, 79.1) P=0.667		
Liston and Dall, 2003 [UK]	• The study describes performance of double reading with	• Program: National Health Service Breast Cancer Screening Programme (NHSBSP)	• Factor of Study: double reading? • Other potential influencing factors:	• **Recall Rate (%)** 3.7 to 6.0% for the five radiologists acting as first readers	**Author Reported Conclusions** • “It is recommended this audit method is adopted by all units in	• Uninformative study • Report of audit results

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
	arbitration but there is no comparison group (e.g., double reading vs. single reading, or blinded vs. non-blinded double reading, or arbitration vs. consensus)	• Study period: 1 April 1995 – 31 March 2002 • Target age: 50-64 • Screening frequency: • Sample size (# of women) = 177,167 • Mean age (years):	**Reading Approach:** films independently read by two readers. In case of disagreement on whether the woman should be returned to routine recall or recalled for assessment, the film was independently reviewed by a third reader. The majority opinion was the basis for action. "it is recognized that the second reader is influenced by the first reader's decision to recall as there is only one set of paper documentation." **Readers' Training:** "five radiologist screen readers of varying experience" **Prior Mammograms:** "All incident screens	• **Cancer detection** 87 (8.1%) of the 1072 cancers were detected following third reader arbitration	the NHSBSP and that the Advisory Committee for Breast Cancer Screening review the policy of single versus double reading. **Author Reported Limitations** • None reported	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			are displayed adjacent to previous screening films to enable comparison." **Technology**: screen film			
Mullen, 2017 [USA]	• N/A	• Program: three outpatient sites of an academic breast imaging • Study period: January 3, 2012 – April 3, 2016. First intervention (awareness): February 3 to September 3, 2015. Second intervention (consensus): September 4, 2015 to April 3, 2016 • Target age: 40 to >65 years • Screening frequency: not reported • Sample size (# of screens = 54,963 (FFDM); 24249 (DBT)	• Factor of Study: consensus double read of potential recalls. **Reading Approach** Baseline: (apparently) single reading Intervention: Consensus double reads of all potential recalls. If a second reader agreed with the recall suggested by the primary reader, the patient was recalled. If the second reader disagreed with the recall, a third reader was asked to provide the decision. All second and third reader reviews and the final report were	• **Recall Rate (%)** FFDM Baseline: 11.1 Consensus: 9.9 P<0.05 DBT Baseline: 7.6 Consensus: 7.2 P>0.05 (not significant) • **Detection Rate (per 1000)** FFDM Baseline: 3.8 Consensus: 5.9 P>0.05 (not significant) DBT Baseline: 4.8 Consensus: 5.7 P>0.05 (not significant) • **PPV1 (%)** FFDM Baseline: 3.4 Consensus: 5.7	• **Author Reported Conclusions** • "…simple interventions, such as personal review of recalls and consensus recall, are associated with decreased recall rates. Consensus recall also increased PPVs for both FFDM and DBT." • "…the reduction in recall rate attributed to review of personal recalls was actually more substantial than the reduction attributed to consensus double reading, for both FFDM and DBT. This unexpected result suggests that the motivated radiologist could invest a small amount of time each week reviewing his or her own recalls, and thereby improve his or	• Intervention study • Information about the Consensus intervention is relevant to this section. See also section "Audit/Performance Feedback" • "An average of 2.3 minutes was used for each recall consultation. The number of consultations per day varied, but ranged between 0 and 10 for the reading radiologist. A third reader was required for arbitration in 2.7% of the consensus recall cases (20 of 728 cases), and time needed for this additional

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Mean age (years): not reported	required within 24 hours of the first reader interpretation to avoid significant delay in releasing the final report to the provider and the patient. **Readers' Training:** radiologists (N=10); all breast imaging specialists with 1 to 22 years of experience (average 10.4 years) **Prior Mammograms:** not reported **Technology:** two dimensional (2D) full-field digital mammography (FFDM) and three-dimensional (3D) digital breast tomosynthesis (DBT) •	P<0.05 DBT Baseline: 6.0 Consensus: 9.0 P<0.05 • "The overall trends of decreased recall rates and increased PPVs were generally distributed across all age groups, although some changes were not statistically significant due to small sample sizes when stratified by age." (see table 2 of the publication)	her personal performance metrics. This result may also suggest that there was marginal remaining opportunity after the awareness phase, therefore decreasing the additional opportunity available for improvement with consensus recall." • "…double-reading only potential recalls was efficient, with an average of 2.3 minutes spent on each case, with an associated decrease in recall rates and increase in PPVs. One must also consider the breast imager's time that is recovered by avoiding the diagnostic workups that result from false-positive screening recalls." • "The effect …appeared greater with FFDM compared to DBT, likely due to the already reduced recall rates associated with the 3D technique. This may also be related to less	consultation was included in the 2.3 minutes per case…"

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					confidence about findings on FFDM, and therefore more opportunity for improvement and better outcomes with two readers. However, both modalities demonstrated a significant increase in PPV1…” **Author Reported Limitations** • “The relatively small sample sizes led to difficulty in detecting significant differences in cancer detection rate when only a few cancers are detected per 1000 screening examinations. With smaller numbers, the cancer detection rate can fluctuate and therefore not reflect the full impact of the interventions.” • “This study was performed at an academic institution with breast imaging specialists, and the techniques may not be effective outside of an	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					academic, subspecialty setting.”	
Posso, 2016 [Spain]	- N/A	- Program: population-based breast cancer screening programme of the Hospital Sant Pau - Study period: June 2009 to May 2013 - Target age: 50-69 years - Screening frequency: - Sample size (# of screens) = 57,157 - Mean age (years):	- Factor of Study: blinded double reading (DR) with consensus or arbitration (C/A) vs. single reading blinded DR with C/A vs. blinded DR without C/A - Other potential influencing factors: **Reading Approach:** blinded double reading with or without consensus or arbitration. The four radiologists were randomly assigned as first or second readers, and then the radiologist who was first reader in one mammogram could be second reader in another mammogram. <u>Double reading with consensus or arbitration</u>. Discordant results	- **False Positive Rate (%)** <u>DR with C/A vs. single reading</u> DR with C/A: 4.5 Single reading: 4.2 P=0.001 <u>DR with C/A vs. DR without C/A</u> DR with C/A: 4.5 DR without C/A: 6.0 P<0.001 - **Detection Rate per 1000** <u>DR with C/A vs. single reading</u> DR with C/A: 4.6 Single reading: 4.2 P=0.283 <u>DR with C/A vs. DR without C/A</u> DR with C/A: 4.6 DR without C/A: 4.7 P=0.986 - **PPV (%)** <u>DR with C/A vs. single reading</u> DR with C/A: 9.3 Single reading: 9.1 P=0.812	- Author Reported Conclusions - “…our results suggest that single reading may have a better interpretative accuracy than double reading. In our study, double reading with consensus and arbitration had more false-positive results than single reading while the positive predictive value was similar in both.” - “Double reading without consensus and arbitration had 1.5 % more false positive results than double reading with consensus and arbitration…Both reading strategies had similar cancer detection rates…” - “…double reading is an expensive strategy that produces more false-positive results than single reading without significantly increasing the cancer detection rate. Our results are not	- Blinded double reading with consensus or arbitration or blinded double reading with unilateral recall.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			were resolved by consensus. When the two readers could not reach consensus, there was arbitration by a third senior radiologist <u>Double reading without consensus or arbitration</u>. Women were recalled if one of the radiologists determined abnormal findings. **Readers' Training: four certified screening** radiologists who read ≥5,000 mammograms per year **Prior Mammograms:** all mammograms were available for comparison at the next screening round.	<u>DR with C/A vs. DR without C/A</u> DR with C/A: 9.3 DR without C/A: 7.1 P=0.001 • **Costs** DR <u>without</u> C/A was 14% more expensive than DR with C/A DR <u>with</u> C/A was 15% more expensive than single reading	conclusive as this study was conducted in a specific context and data regarding interval cancers was not available." • Author Reported Limitations • "…it is difficult to transfer costs of reading strategies from one country to another because breast cancer screening programmes differ considerably in unitary costs, working times, and protocol-reading variables." • "we did not analyze indirect and non-health related costs because we performed the analysis from the health system perspective." • "this study was conducted in a very specific context: four certified, highly trained, radiologists were involved in reading the mammograms. Therefore, the generalization of our results to other contexts with less trained radiologists is	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Technology: digital		doubtful and additional data are needed." • "…we were unable to calculate sensitivity and specificity of the programme because we cannot obtain information about interval cancers as data from population-based cancer registries are not available in our city."	
Salas, 2011 [Spain]	• N/A	• Program: Eight regional breast cancer screening programs covering 44% of the target population participated in this study. • Study period: 1990-2006 • Target age: four of the eight programs start screening at age 45 and the remaining four start at age 50 years • Screening frequency: every 2 years	• Factor of Study: single vs. double reading • Other potential influencing factors: **Reading Approach:** not described **Readers' Training:** radiologists **Prior Mammograms:** not reported **Technology:** screen-film and digital (mammographic technique was one of the variables	• **Odds ratio (OR) for the false positives risk (95% CI)** <u>False positives, any procedure (FP):</u> Single reading: 1.00 (ref.) Double reading: 1.36 (1.23, 1.51) <u>False positives, invasive procedure (FPI):</u> Single reading: 1.00 (ref.) Double reading: 1.04 (0.67, 1.62)	**Author Reported Conclusions** • "Programme-related variables such as double reading increase the FP risk…" **Author Reported Limitations** • None reported	• The focus of this study was the age at which breast cancer screening starts. Screen reading strategy was one of the variables included in the analysis. • Double reading approach (e.g., blinded or non-blinded, consensus/ arbitration or unilateral recall) is not described.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Sample size (# of mammograms) = 4,739,498; (# of women) = 1,565,364 • Mean age (years):	included in the analysis) •			
Shaw, 2009 [Ireland]	• N/A	• Program: the Irish National Breast Cancer Screening Program (NBSP); a screening center serving the eastern part of Ireland • Study period: 2000-2005 • Target age: 50-64 • Screening frequency: biennial • Sample size (# of screens) = 128,569 • Mean age (years): 57.2±4.2 *SD)	• Factor of Study: consensus review of discordant findings in double reading • Other potential influencing factors: **Reading Approach:** Independent reading by two radiologists. A consensus panel met twice a week. The panel consisted of 3 to 5 radiologists and usually included one or both original researchers. A woman was recalled if any member of the panel recommended referral. **Readers' Training:** Five radiologists trained in screening and diagnostic mammography;	• **Recall strategies used for comparison:** Highest reader recall: a patient is recalled if her findings are deemed abnormal by either reader Unanimous recall: none of the patients with discordant findings was referred for further assessment • **Recall Rate (%)** Consensus review vs. Highest reader recall: 4.41 vs. 4.97 [relative increase of 12.69% with highest reader recall] Consensus review vs. Unanimous recall: 4.41 vs. 3.94 [relative decrease of 10.66% with unanimous recall If all patients with discordant calcifications were recalled,	**Author Reported Conclusions** • "Consensus review of cases with discordant findings improves cancer detection (consensus review led to the identification of 7.3% of all cancers diagnosed at our center between 2000 and 2005) and maintains a low false-negative rate (0.72% of all cancers)." • "Use of the highest reader recall method, in which a patient is recalled if her findings are deemed abnormal by either reader, could potentially increase the cancer detection rate by 0.6 per 1000 women screened but would increase the recall rate by 12.69% and the number of false-positive findings by 15.37%."	• Independent double reading • The strategy for resolution of discordant opinions is different from that described in other studies (see "Reading Approach"

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			three of them had 3-10 years of screening experience at the start of study and two had just completed fellowship training. All radiologists read >5000 mammograms per year. **Prior Mammograms:** comparison of old and new mammograms is listed as one of the practices adopted by the Irish National Breast Cancer Screening Program (NBSP) **Technology:** digital	the overall recall rate would increase by 0.05%. - **Detection Rate per 1000** Consensus review vs. Highest reader recall: 7.47 vs. 7.53 Consensus review vs. unanimous recall: not reported	- "Consensus review facilitates the early diagnosis of cancers that tend to exhibit subtle findings at mammography." - "Consensus review of discordant findings substantially reduces the number of normal cases recalled for assessment…" **Author Reported Limitations** - "A limitation of our study was our use of a nonuniform review panel. Membership was subject to change from week to week and depended on the number of radiologists available to participate in the discussion. Members with different levels of experience reviewed the cases, and the panel usually included one or both of the original readers. While these factors introduced bias into the data set, it would be impossible to implement a uniform	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					review process in actual practice.”	
Taylor-Phillips, 2016 [UK]	• N/A	• Program: a multi-center, double-blind, cluster randomized clinical trial that included 46 breast screening centers from the NHSBSP in England • Study period: December 20, 2012 to November 3, 2014 • Target age: 50-70 years • Screening frequency: every 3 years • Sample size (# of women) = 1,194,147 • Mean age (years): 59.3 (SD, 7.49) •	• Factor of Study: changing order in which the two readers in a double reading program examine a batch of mammograms • Intervention study: the two readers examined each batch of mammograms in the same order (control group) or in the opposite order to one another (intervention group). • Other potential influencing factors: **Reading Approach:** Batch reading (each batch included ≈40 mammograms from a single mammography machine in a single day). Readers were encouraged to read the batches independently but	• **Recall Rate (%)** Intervention group: 4.14 Control group: 4.17 Difference: -0.03 (95% CI: -0.10, 0.04) “…recall rate for individual readers (the proportion of women that 1 reader determined should be recalled) reduced with time on task. The odds of recall decreased over the course of examining 40 cases (OR,0.83; 95%CI,0.81-0.85). The reduction was similar in the model adjusted for woman’s age and previous attendance (OR, 0.89; 95%CI, 0.87-0.91…)” • **Detection Rate (%)** Intervention group: 0.88 Control group: 0.87 Difference: 0.01 (95% CI: -0.02, 0.04) “…cancer detection rate for individual readers did not	**Author Reported Conclusions** • “The intervention did not influence cancer detection rate, recall rate, or rate of disagreement between readers. There was no pattern of decreasing cancer detection rate with time on task as predicted by previous research on vigilance decrements as a psychological phenomenon. Instead there was a gradual decrease in recall rate, with an increase in PPV and a decrease in false-positive recall of women with time on task. This may reinforce and explain previous observational research that identifies that recall rate is reduced when grouping women’s cases into batches.” **Author Reported Limitations**	• The aim was to determine whether a vigilance decrement (reduced detection rate with time on task) exists in breast cancer screening and whether changing the order in which two readers examine a batch of mammograms can increase the cancer detection rate [Assuming that the two readers experience peak vigilance at different points within the reading batch.] • The authors explain the contradiction between their results and the results of previous studies demonstrating

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			can access the other reader's decision. In 16 of the 46 centers, reader 2 was blinded to the decision of reader 2. All centers used arbitration in case of disagreement: 13 used a single third reader and 33 used a group consensus of two or more readers. **Readers' Training:** 186 radiologists, 143 radiography advanced practitioners, 31 breast clinicians. All readers were accredited by the NHSBSP, read ≥5000 cases per year, participated in assessment clinics and regularly audited their performance.	change with time spent on task, as represented by near identical odds of detecting cancer between the first and 40th case (OR, 0.987; 95% CI, 0.929-1.048). Results were very similar in the model adjusted for the characteristics of the woman screened (OR, 0.995; 95% CI, 0.938-1.055…)" • **Rate of Disagreement (%)** Intervention group: 3.43 Control group: 3.48 Difference: -0.05 (95% CI: -0.11, 0.02)	• "…we did not control for or measure working conditions, some of which may affect whether there is a vigilance decrement." • "…we did not specify or measure the length of each reader's work week, the proportion of his/her time spent working in breast screening or reading mammograms, the number of work hours or type of work activities each day, number of breaks taken, or self-perceptions of fatigue." • Center-level variation in management of individual performance were not recorded. • "…the trial did not attempt to implement blinding of reader 2 to the decision of reader 1 in centers in which this was not standard practice, as limiting reader's access to computerized and paper notes was not considered possible	vigilance decrement by the fact that most previous studies were conducted in psychology laboratories rather than in real-life settings.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Prior Mammograms: not reported **Technology:** digital •		without compromising patient safety.”	
Audit/Performance Feedback						
Carney, 2011 [USA]	• N/A	• Program: four mammography registries contributing to BCSC • Study period: unclear; data on radiologists’ recall rates were collected for 2003-2004; to be eligible, radiologists had to actively interpret mammograms between January 2006 and September 2007, • Target age: not reported • Screening frequency: not reported • Sample size (# of **radiologists**) • N=196 eligible (actively interpreting	• Factor of Study: performance feedback; educational intervention. Interactive web-based intervention included three components (modules): 1) Peer comparison audit data on performance indicators; the first module was also aimed at explaining audit statistics and how they were derived. 2) Addressing radiologists’ misconception about women’s’ risk of breast cancer. 3) Addressing radiologists’	• **Self-reported recall rates (%)** Early intervention group: 11.8 Late intervention (control) group: 18.2 P=0.015 • 95% found the program moderately to very helpful in understanding how basic performance measures are calculated • 93% found viewing their performance measures moderately to very helpful • 83% found it was moderately to very helpful to know that the breast cancer risk in their screening population was lower than they perceived • The percentage of radiologists who reported that the risk of medical malpractice influenced their recall rates: pre-intervention	**Author Reported Conclusions** • “…radiologists who begin an internet-based tailored intervention designed to help reduce unnecessary recall in mammography will likely complete it, though only about half who consented to the study actually completed the intervention. Greater than 90% of participants found the intervention useful in helping them understand why their recall rates may be elevated. More research needs to be done to understand how best to engage radiologists in undertaking educational programs on the internet.”	• Intervention study • In this article, only self-reported recall rates are reported. See Carney et al. 2012 (below) for actual recall rates

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		mammograms in 2006-2007) and invited to participate N=74 agreed to participate N=40 randomized to early intervention group N= 34 randomized to late intervention group (served as a control group for a 9-month follow-up period, after which they were invited to receive the educational intervention) N=41 completed the intervention (27 early intervention; 14 late intervention) N=5 started but did not complete the intervention N=28 never started the intervention • Mean age (years): not reported	misconceptions regarding malpractice related to breast imaging. Radiologists could click on links embedded in the intervention to read relevant literature. • Other potential influencing factors: **Reading Approach:** not reported **Readers' Training:** radiologists **Prior Mammograms:** not reported • **Technology**: not reported	36.3%; post-intervention 17.8% • The percentage of radiologists who reported that the risk of medical malpractice influenced their recommendation for breast biopsy: pre-intervention 36.4%; post-intervention 17.3% • >75% of radiologists correctly answered post-intervention knowledge questions	**Author Reported Limitations** • "…this may not be a representative sample of radiologists across the United States." • "The findings could also be affected by selection bias, though our assessment of the characteristics of those who did and did not consent was reassuring." • "Though the intervention was designed to address adult learning principles, some aspects of adult learning theory could not be accommodated in our intervention." • "We also have yet to conduct an analysis of the impact of this intervention on clinical performance…"	
Carney, 2012 **[USA]**	• N/A	• See Carney et al. 2011 (above) • This study included radiologists for whom screening mammography	• See Carney et al. 2011 (above)	• **Recall Rate (%)** ***Analysis 1: Radiologists who consented to the intervention but <u>did not</u> complete it***	**Author Reported Conclusions** • "In conclusion, we developed and implemented an innovative, web-based	• Intervention study • <u>CME</u>: continuing medical education • "There were no statistical differences in

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		interpretation data were available: 1) 32 radiologists who consented to the intervention and completed it (23 assigned to the intervention group and 9 assigned to the control group); 2) 22 radiologists who consented to the intervention but did not complete it (10 assigned to the intervention group and 12 assigned to the control group)		<u>Assigned to the intervention group (n=10)</u> Baseline (9 months prior to consent): 11.0 0-9 months after consent (T1): 9.4 9-18 months after consent (T2): 9.7 <u>Assigned to the control group (n=12)</u> Baseline (9 months prior to consent): 9.6 0-9 months after consent (T1): 9.0 9-18 months after consent (T2): 9.3 ***Analysis 2: Radiologists who consented to the intervention and completed it*** <u>Intervention group (n=23)</u> Baseline (9 months prior to consent): 11.2 0-9 months after completion of the CME (T1): 10.8 9-18 months after completion of the CME (T2): 10.4	educational program that take advantage of computerized registry data collected on community radiologists to provide them with individualized audit feedback. Our study resulted in a null effect, which may indicate a single intervention is not adequate to change excessive recall among radiologists who undertook the intervention we were testing. It is likely that more complex approaches are needed to change radiologists practice patterns." **Author Reported Limitations** • "small sample size, which affected our ability to power this study to detect meaningful differences in recall" • "…more work is needed to understand how best to influence radiologists practice."	radiologists' characteristics according to study group assignment among those who consented and completed and those who consented but did not complete the intervention." • Performance measures other than recall rates are not reported.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				Control group (n=9) Baseline (9 months prior to consent): 8.7 0-9 months after consent (T1): 8.8 0-9 months after completion (T2): 9.2 • **Probability of being recalled; adjusted for mammography registry, patients' and radiologists' characteristics** Radiologists who completed the intervention (n=22) in the Intervention Group vs. all radiologists (n=19) in the Control Group Relative change from baseline to T1: OR=1.11 (95% CI: 1.00, 1.23) Relative change from prior to consent to T2: OR=1.09 (95% CI: 0.98, 1.21) Radiologists who completed the intervention in the Intervention Group (n=22) vs. radiologists who completed the intervention in the Control Group (n=9)		

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				Relative change from baseline to T1: OR=1.12 (95% CI: 1.00, 1.27) Relative change from prior to consent to T2: OR=1.10 (95% CI: 0.96, 1.25) Radiologists in the Control Group who completed the late intervention vs. those who did not Relative change from baseline to T1: OR=0.97 (95% CI: 0.83, 1.14) Relative change from prior to consent to T2: OR=0.98 (95% CI: 0.83, 1.16)		
Geertse, 2015 [The Netherlands]	•	• Program: this article reports on the results of four triennial audits of all 17 Dutch reading units performed by the Dutch Reference Center for Screening (LRCB) • Audit series: 1996-2000, 2001-2005, 2003-2007, 2010-2013	• Factor of Study: audit/performance feedback • The performance of the team of radiologists of a reading unit (RU) and not an individual performance is assessed. • The audit follows a fixed protocol and includes two parts: evaluation of screening outcomes and radiological	• **Recall Rate (%)** 1990-1997: 0.66 (0.5 - 1.0) 1998-2003: 1.07 (0.7 - 1.5) 2001-2006: 1.22 (0.7 - 1.9) 2006-2009: 1.58 (1.0 - 2.2) Trend: +0.29 (95% CI: 0.23, 0.35) P=0.000 • **Detection Rate (per 1000 screened)** 1990-1997: 3.3 (2.7 - 4.1) 1998-2003: 4.5 (3.9 - 6.4) 2001-2006: 4.8 (3.5 - 5.6) 2006-2009: 5.4 (4.3 - 6.2)	**Author Reported Conclusions** • "Over the four audit series we observed a positive trend in recall rate, detection rate and sensitivity…" • The recall rate in the DBCSP is one of the lowest worldwide. Low recall rates may result in more missed subtle cancers." • "During the audits, the LRCB advised all RU radiologists to lower	• Because many changes occurred during the study period (switch from SFM to DM, from non-blinded to blinded double reading, from single-view to two-view mammography), it is unclear whether the observed trends in performance indicators were

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Outcome data: 1990-1997; 1998-2003; 2001-2006; 2006-2011 • Target age: 50-74 years • Screening frequency: every 2 years • Sample size (# of women): 966,573; 1,913,739; 2,053,796; 3,106,806 at the four series, respectively. • Mean age (years): not reported •	review of mammograms • Starting from 2010, the radiological review includes 40 interval cancers, 40 stage II cancers and 40 consecutive recall cases. • The review by the radiologists from the audit team takes place in the presence of RU radiologists, and there is an open discussion. • Feedback was provided directly following the audit at a final meeting. • A report summarizing results and providing recommendations was prepared. • After the audit, the LRCB organized a 2-hour refresher course on site to provide feedback to radiologists who could not attend the audit.	Trend: +0.6 (95% CI: 0.5, 0.8) P=0.000 • **PPV of recall (%)** 1990-1997: 51.9 (30.0 - 66.7) 1998-2003: 43.5 (32.9 - 60.3) 2001-2006: 41.5 (27.3 - 59.1) 2006-2009: 35.5 (26.7 - 47.1) Trend: −5.2 (−3.8, −6.6)0 P=0.000 **Sensitivity (%); data from the Thirteens Evaluation Report of the National Evaluation Team for Breast Cancer** 1990-1997: 64.6 1998-2003:68.7 2001-2006: 70.5 2006-2009: 71.6 Trend: +2.3 (0.2, 4.4) P=0.043	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." • "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	associated with the audit/performance feedback.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			• Other potential influencing factors: **Reading Approach:** non-blinded double reading in the period of screen-film mammography (SFM); blinded double reading after the conversion to digital mammography (DM) in 2008-2010. Discrepant findings were resolved by consensus between two readers or arbitration by a third reader **Readers' Training:** radiologists **Prior Mammograms:** not reported **Technology:** SFM; conversion to DM in 2008-2010. During the study period, the units switched from		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." **Author Reported 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 affect for instance recall rate.... In our study, we were not able to investigate the	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			single-view to two-view mammography •		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.”	
Hofvind, 2016 [member countries of the International Cancer Screening Network (ICSN)]	• This article describes audit feedback but does not provide information regarding its possible influence on program performance indicators	• Program: a web-based survey included 17 screening programs in member countries of the International Cancer Screening Network (ICSN) • Study period: 2012 • Target age: varies by country/screening program (see table 1 of the publication) • Screening frequency: varies by country/screening	• Factor of Study: audit feedback • Other potential influencing factors: **Reading Approach (reported):** independent double reading (7 programs – Australia, France, Luxembourg, the Netherlands, Norway, Switzerland, UK); double reading (2 programs – Sweden, Japan); independent double or double (1	• Audit feedback was directed to the individual reader and/or the facility • Purposes of the reader and facility audit feedback listed by the participants: identify result outliers; monitor performance for quality assurance; identify readers or facilities that need special training; compare between readers; document the results (see table 3 of the publication) • Target audience: the readers in 13 of 14 programs that responded to the question; facilities and health administrators in Australia, Catalonia, Franc and	**Author Reported Conclusions** • The purpose, target audience, performance measures included, form and frequency of the audit feedback varied amongst mammographic screening programmes. These variations may provide a basis for those developing and improving such programmes.” **Author Reported Limitations** • “…responses were received from less than	• Numbers reported only for programs that responded to each question.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		program; every 2 years in most programs (see table 1 of the publication) • Sample size (# of women screened in 2010): from 1459 (Luxembourg) to 2,492,863 (Japan) • Mean age (years): not reported	program Catalonia); independent double reading with or without CAD/other (1 program – Denmark); single reading only (2 programs -Navarra, Ontario); a mix of different reading procedures (3 programs – Saskatchewan, Quebec, the US). **Readers' Training:** "All programmes had recommendations or requirements for individuals to be eligible both to start and/or to continue reading screening mammograms." (see table 2 of the publication) **Prior Mammograms:** not reported	Luxembourg; facility in Norway and Saskatchewan • The main target audience for facility-level feedback: readers (8 programs); facilities (7 programs); health administrators (4 programs) • Responsible for running the analyses for readers: readers (2 programs); analyses on the local level (6 programs); analyses on the regional level (2 programs); analyses on the national level (1 program) Responsible for running the analyses for facilities: analyses on the regional (3 programs); on the national level (3 programs); independent units (3 programs); medical leader (3 programs); other (1 program) • Reader level feedback reports included: screening volume and recall rates (14 programs); screen-detected cancer (13 programs); interval cancer rate (8 programs); characteristics of the interval tumors (3 programs) • Facility-level feedback reports included: recall rate (11 programs); screening	half of ICSN member countries." • "We do not know if the low response rate was a result of not having audit feedback or not being able to identify an appropriate person to complete the survey." • "Having only one person representing a programme respond to the survey might also bias the results." • "…parts of the survey were not completed, we cannot tell whether this is because the programme does not provide that kind of audit feedback, if data were not available, or if that part was simply not filled in."	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Technology: not reported	volume and rate of screen-detected cancer (10 programs); interval cancer rate (6 programs); PPV (7 programs); histologic characteristics of screen detected cancers (9 programs) and interval cancers (5 programs) • Frequency of individual audit feedback: annually or more frequently (10 programs); ad hoc (4 programs); web-based data accessible all the time (1 program – US) • Frequency of facility-level audit feedback: annual (7 programs); ad hoc (Norway); “infrequently” (Luxembourg); “other” (the Netherlands) • Actions if guidelines/benchmarks were not achieved: remedial support/training for readers (4 programs); remedial training for the facility (1 program); no action (3 programs); removed readers from the program (2 programs); “provided more skills” (1 program); meetings to discuss actions (1 program)		

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				• Actions regarding the facility if guidelines/benchmarks were not achieved: remedial training (1 program); more frequent monitoring (2 programs); further investigation (7 programs); "provided more skills (1 program); "identification of facilities that need intervention with a subsequent decision on the action" (1 program)		
Liston and Dall, 2003 [UK]	• The study reports on audit results (performance of five radiologists) but there is no analysis of the influence of the audit on the performance of these radiologists	• Program: National Health Service Breast Cancer Screening Programme (NHSBSP) • Study period: 1 April 1995 – 31 March 2002 • Target age: 50-64 • Screening frequency: • Sample size (# of women) = 177,167 • Mean age (years):	• Factor of Study: audit • Other potential influencing factors: **Reading Approach:** films independently read by two readers. In case of disagreement on whether the woman should be returned to routine recall or recalled for assessment, the film was independently reviewed by a third reader. The majority opinion was the basis for action. "it is	• **Recall Rate (%)** 3.7 to 6.0% for the five radiologists acting as first readers • **Cancer detection** • 87 (8.1%) of the 1072 cancers were detected following third reader arbitration	**Author Reported Conclusions** • "It is recommended this audit method is adopted by all units in the NHSBSP and that the Advisory Committee for Breast Cancer Screening review the policy of single versus double reading. **Author Reported Limitations** • None reported	• Uninformative study • Report of audit results

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			recognized that the second reader is influenced by the first reader's decision to recall as there is only one set of paper documentation." **Readers' Training:** "five radiologist screen readers of varying experience" **Prior Mammograms:** "All incident screens are displayed adjacent to previous screening films to enable comparison." • **Technology**: screen film			
Mullen, 2017 [USA]	• N/A	• Program: three outpatient sites of an academic breast imaging • Study period: January 3, 2012 – April 3, 2016. First intervention	• Factor of Study: performance feedback (awareness). Intervention. Phase 1: Each radiologist compared his/her individual performance to that	• **Recall Rate (%)** FFDM Baseline: 11.1 Awareness: 9.2 P<0.05 DBT Baseline: 7.6 Awareness: 6.6	**Author Reported Conclusions** • "We have shown that simple interventions, such as personal review of recalls …are associated with decreased recall rates."	• Intervention study • Information about the Awareness intervention is relevant to this section. See also "Double reading"

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		(awareness): February 3 to September 3, 2015. Second intervention (consensus): September 4, 2015 to April 3, 2016 • Target age: 40 to >65 years • Screening frequency: not reported • Sample size (# of screens = 54,963 (FFDM); 24249 (DBT) • Mean age (years): not reported	of the group. The group discussed perceptions of recall/performance, e.g., most frequent reasons for recall, individual fears prompting recall. A goal was set to reduce the division's recall rate to 5%, while monitoring cancer detection rate and PPV Phase 2: Each radiologist weekly reviewed the imaging and reports of his/her recalls, and then the imaging and reports from the subsequent diagnostic evaluation/biopsy for each recalled patient. • Other potential influencing factors: **Reading Approach:** single reading for the	P<0.05 • **Detection Rate (per 1000)** FFDM Baseline: 3.8 Awareness: 3.1 P>0.05 (not significant) DBT Baseline: 4.8 Awareness: 6.2 P>0.05 (not significant) • **PPV1 (%)** FFDM Baseline: 3.4% Awareness: 3.1% P>0.05 (not significant) DBT Baseline: 6.0 Awareness: 10.8 P<0.05 • "The overall trends of decreased recall rates and increased PPVs were generally distributed across all age groups, although some changes were not statistically significant due to small sample sizes when stratified by age." (see table 2 of the publication)	**Author Reported Limitations** • "The relatively small sample sizes led to difficulty in detecting significant differences in cancer detection rate when only a few cancers are detected per 1000 screening examinations. With smaller numbers, the cancer detection rate can fluctuate and therefore not reflect the full impact of the interventions." • "This study was performed at an academic institution with breast imaging specialists, and the techniques may not be effective outside of an academic, subspecialty setting."	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			awareness intervention **Readers' Training:** radiologists (N=10); all breast imaging specialists with 1 to 22 years of experience (average 10.4 years) **Prior Mammograms:** not reported **Technology:** two dimensional (2D) full-field digital mammography (FFDM) and three-dimensional (3D) digital breast tomosynthesis (DBT)			
Comparison with Prior Mammograms						
Hayward, 2016 [USA]	•	• Program: • Study period: 14 June 2010 – 3 March 2015 • Target age: not reported	• Factor of Study: two or more prior mammograms vs. a single prior mammogram • Other potential influencing factors:	• **Recall Rate (%)** No priors: 16.6 (includes prevalent screens) One prior: 7.8% Two or more priors: 6.3% • **Recall Rate, unadjusted odds ratio [OR (95% CI)]**	**Author Reported Conclusions** • "…we have shown that the screening mammography recall rate decreases while the PPV1 and CDR increase when two or	• The incremental efficacy of two or more prior mammograms was tested relative to a single prior mammogram because comparisons to

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Screening frequency: not reported • Sample size (# of screens) = 46,288; (# of women) = 22,792 • Mean age (years): 59±12.2	**Reading Approach:** not reported, presumably single reading **Readers' Training:** radiologists **Prior Mammograms:** factor of study • **Technology**: only digital mammography during the study period; however, prior examinations could be screen film. It is not clear whether prior screen films were digitalized.	1 vs. 0: 0.430 (0.379, 0.489). P<0.0001 ≥2 vs. 0: 0.340 (0.309, 0.374). P<0.0001 ≥2 vs. 1: 0.789 (0.711, 0.877). P<0.0001 • **Recall Rate, odds ratio adjusted for age [OR (95% CI)]** 1 vs. 0: 0.470 (0.413, 0.536). P<0.0001 ≥2 vs. 0: 0.406 (0.366, 0.451). P<0.0001 ≥2 vs. 1: 0.864 (0.776, 0.962). P=0.0074 • **Cancer Detection Rate (per 1000) (95% CI)** 1 mammogram: 4.3 (2.8, 6.4) ≥ 2 mammograms: 6.6 (5.8, 7.5) Combined: 6.3 (5.6, 7.1) • **PPV (%)** 1 mammogram: 0.056 (0.035, 0.077) ≥ 2 mammograms: 0.105 (0.093, 0.118) Combined: 0.097 (0.086, 0.011)	more prior examinations are used for comparison relative to comparison with a single prior examination. Our findings suggest that, at screening mammography, radiologists who compare with more than a single prior examination will have more true–positive and fewer false–positive outcomes." **Author Reported Limitations** • "… retrospective design, which may lead to selection bias" • "…we were only able to obtain information regarding cancer diagnoses from our institutional pathology database as we did not have linkage to a tumor registry for the study period." • "Another confounding factor in our study is the fact that the median time interval to the last prior comparison	women with no prior mammograms was confounded by prevalent screens.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					mammogram was 12 months longer in the single prior group as compared to the multiple prior group. Therefore, we cannot exclude the possibility that the availability of more recent priors may have contributed to the recall rate reduction in the multiple prior group." • "…we did not differentiate between the number of comparison exams within the multiple prior group in our analysis. We sought only to establish that comparing with multiple priors is better than comparing to a single prior."	
Klompenhouwer, 2014 [The Netherlands]	• N/A	• Program: the southern screening mammography region of the Netherlands • Study period: January 2000 – July 2011. SFM screens	• Factor of Study: "influence of comparison with scanned in priors instead of hard copy priors at FFDM on the proportion of women recalled twice for the same	• **Women recalled twice for the same lesion (% of all recalls)** SFM: 37 of 4,140; 0.9% of recalls FFDM: 52 of 2,782 (1.9% of all recalls) P<0.001	**Author Reported Conclusions** • "During the first screening round at FFDM, a significantly larger proportion of recalls included women who had been recalled twice for the same	• These findings are important because the rate of re-attendance at screening was significantly lower for women who had had a repeated false positive recall,

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		were performed between January 2000 and April 2010; FFDM screens were performed between May 2009 and July 2011 • Target age: 50-75 • Screening frequency: biennial • Sample size (# of screens) = 302,912 SFM screens and 90,288 FFDM screens • Mean age (years):	mammographic abnormality” • Other potential influencing factors: **Reading Approach:** double reading **Readers’ Training:** 12 certified screening radiologists; each evaluated ≥3,000 screening mammograms per year (mean 6,000) **Prior Mammograms:** always available at the subsequent screening round with SFM; the most recent screen-film mammograms were digitalized • **Technology**: during the study period, the program switched from screen-film to full-field digital mammography	• **Malignancies detected at the second recall (% of all screen detected cancers)** SFM: 13 (0.8% of screen detected cancers) FFDM: 8 (1.3% of all screen detected cancers) • **PPV of second recall (%)** SFM: 35.1% FFDM: 15.4% P=0.03 • Blinded review showed that, if a hard copy SFM examination, in addition to the most recent digitalized SFM screen, were available at the time of FFDM screening, the number of second recalls at FFDM would have been 32 instead of 52 (39.5% reduction); none of the 20 women who would not have been recalled were ultimately diagnosed with breast cancer. PPV of second recall would have been 25.0%	lesion and breast cancer was significantly less often diagnosed in these women than at SFM, with a concomitant lower PPV of recall. However, the availability of the hard copy SFM screen, in addition to the digitized SFM screen, would have reduced the number of repeatedly recalled women at FFDM by almost 40 %.” • ““This observation underscores the importance of having prior screens available for comparison. Some mammographic abnormalities may come and go, so the availability of older images for comparison is crucial and can lower the recall rate. Failure to do so will increase the amount of false positive screens and eventually lower the re-attendance rate and the effectiveness of screening mammography.”	especially if both recalls were for the same mammographic lesion. • Re-attendance was 93.2 % for women with a negative screen, 65.4% for women recalled once, 56.7% for women recalled twice for different lesions and 44.3% for women recalled twice for the same lesion.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
					Author Reported Limitations • Although all radiologists had more than 5 years of experience with working in a digital radiology environment, including digital mammography, none of them had experience with the use of FFDM in a screening setting at the start of FFDM. screening. However, it is unlikely that our results have been influenced by a learning effect, as the recall rate, cancer detection rate and PPV of recall did not change during the digital screening period…” • “…the FFDM group was restricted to women who were digitally screened for the first time and we cannot predict the long-term impact on second recalls for the same lesion at successive digital screening rounds.”	

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
Taylor-Phillips, 2012 [UK]	• N/A	• Program: a test set was assembled from a UK breast screening center • Study period: cases detected between March 2005 and June 2007 were included • Target age: 50-70 years • Screening frequency: every 3 years • Sample size (# of screens) = 160 anonymized cases (66 incident round cancers detected at digital screening and 94 benign/normal cases). Benign/normal cases were randomly selected from a database of difficult benign/normal cases. • Mean age (years): not reported	• Factor of Study: comparison with prior mammograms • Other potential influencing factors: **Reading Approach:** double reading with arbitration was used in the center from which the test set was assembled. The data for false positive and recall rates were from single readers in the study; an analysis was conducted to convert the results from single reader to double reader with arbitration. **Readers' Training:** four radiologists and four radiography advanced practitioners trained to read mammograms; 3-14 years of experience reading	• Cancer detection rate was greater when using prior mammograms in either format compared to using no prior mammograms. There was no difference between using film or digitalized prior mammograms. • The number of false positive cases without prior mammograms was 24% higher than with film prior mammograms (p=0.03) and 28% higher than with digitalized priors (P<0.05). The difference between using film and digitalized prior mammogram display was not significant (p=0.9). • Overall (combined results for digitalized and film prior mammograms): 26% increase in false positives when prior mammograms were not used relative to when they were used in either format (p=0.02). This would correspond to an increase in recall rate at the study hospital from 4.3% to 5.5% with no increase in cancer detection rate. • Estimated cost of this increase: 13,666 euros per 10,000 screened; estimated	**Author Reported Conclusions** • "In the transition to digital mammography screening prior mammograms should be displayed, as using film or digitised prior mammograms was found to improve performance. If the results translated into everyday screening, then a decision not to use the prior mammograms may increase the recall rate at the study hospital from 4.3% to 5.4% with no change in cancer detection rate. The cost associated with the equipment and staff to display the prior mammograms would be offset by avoiding the high cost of the extra unnecessary recalls." **Author Reported Limitations** • "Participants in our study were aware that they were reading a difficult case set, and	• Test set • "The same 160 cases were each read three times on a digital workstation: with film prior mammograms; digitised prior mammograms; and without prior mammograms. At least one month elapsed between participants re-reading the same cases. Reading sessions were undertaken by each participant on the same day of the week and at the same time of day to reduce confounding due to location, situation or timing. Each session involved reading no more than 54 cases to reduce the effects of fatigue."

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			mammograms (mean 7 years) **Prior Mammograms:** all cases had film prior mammograms from three years previously • **Technology**: digital mammography; prior mammograms were film or digitalized film	cost of digitalized of film display of prior mammograms: 13,115 and 7,612 euros per 10,000 screens, respectively.	that their performance was being measured. This may have resulted in greater vigilance than in a real-life screening situation where a high volume of cases is read in a short space of time.” • “The data presented here are from eight participants all of whom work at the same breast screening centre. A greater number of participants from a wider range of screening centres would have increased generalisability…” • “the increased prevalence of abnormal cases in this study may have led to an underestimate of the number of cases which would be recalled in screening practice.”	
Yankaskas, 2011 [USA]	• N/A	• Program: facilities participating in the Carolina Mammography Registry	• Factor of Study: comparison with prior mammograms • Other potential influencing factors:	• **Recall Rate (%)** Comparison mammogram No: 14.9 Yes: 6.9 Change on comparison mammogram	**Author Reported Conclusions** • “…having comparisons mammograms in a large community-based population leads to	• The meaning of change Is unclear. In some cases, the authors refer to “change on comparison mammograms” or

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		• Study period: 1994-2008 • Target age: ≥40 years • Screening frequency: • Sample size (# of screens) = 1,157,980; (# of women) = 435,183 • Mean age (years):	**Reading Approach:** not reported **Readers' Training:** **Prior Mammograms:** factor of study • **Technology**: not reported	No: 2.0 Yes: 41.1 • **Cancer Detection Rate (per 1000)** Comparison mammogram No: 7.1 Yes: 3.7 Change on comparison mammogram No: 0.8 Yes: 25.4 • **PPV (%)** Comparison mammogram No: 4.8 Yes: 5.4 Change on comparison mammogram No: 3.9 Yes: 6.0	lower recall rates and higher overall specificity." • "…comparison mammograms lead to lower sensitivity." • "Comparison mammograms are reviewed to look for change, and whether or not change is noted has a large effect on the recall rates and performance measures. Recall rates were 2.2 times higher when change was noted compared with when no change was noted." **Author Reported Limitations** • "The group of women in whom no comparison mammograms were available (6.9% of subjects) …had a higher proportion of prevalent screening mammograms. As a result, they had more cancers, a higher sensitivity, and a higher cancer detection rate."	"change(s) in the comparison image", in other cases (including the title) they refer to "change from the comparison mammogram". • Comparison of recall and cancer detection rates in women with and without prior mammogram may not be meaningful because higher proportion of screens for which no prior mammograms are available are prevalent screens. and this does not appear to be accounted for in the analyses.

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
Mammographic Compression						
Holland, 2016 [The Netherlands]	- N/A	- Program: the Dutch breast cancer screening program - Study period: 2003-2011 - Target age: 50-75 years - Screening frequency: - Sample size (# of screens) = 113,464 - Mean age (years):	- Factor of Study: compression pressure applied during the acquisition of the mammogram - Other potential influencing factors: **Reading Approach:** not reported **Readers' Training:** not reported **Prior Mammograms:** not reported - **Technology**: digital	- **Recall Rate per 1000 (95% CI)** ≤7.68 kPa: 21.9 (20.0–23.8) >7.68, ≤9.18 kPa: 20.9 (19.0–22.8) >9.18, ≤10.71 kPa: 21.8 (19.9–23.7) >10.71, ≤12.81 kPa: 20.9 (19.0–22.8) >12.81 kPa: 22.1 (20.1–24.0) Pearson's χ^2: 0.858 - **False Positive Rate per 1000 (95% CI)** ≤7.68 kPa: 16.3 (14.7–18.0) >7.68, ≤9.18 kPa: 14.4 (12.8–15.9) >9.18, ≤10.71 kPa: 14.6 (13.1–16.2) >10.71, ≤12.81 kPa: 15.5 (13.9–17.1) >12.81 kPa: 17.2 (15.5–18.9) Pearson's χ^2: 0.088 - **Cancer Detection Rate per 1000 (95% CI)** ≤7.68 kPa: 5.5 (4.6-6.5) >7.68, ≤9.18 kPa: 6.5 (5.5–7.6) >9.18, ≤10.71 kPa: 7.1 (6.0–8.2) >10.71, ≤12.81 kPa: 5.4 (4.4–6.3)	**Author Reported Conclusions** - "Significant differences across the five groups are seen for the PPV and the cancer rate. Here the highest PPV is observed in group 3. No statistically significant differences were found in the recall rate and the false positive rate. Even though differences are not significant, there is a trend that the groups with a moderate pressure have lower false positive rate compared to the first and last groups. And that the highest pressure reduces the cancer detection rate." - "In this study, differences in performance measures were observed with respect to different pressure categories. Although only PPV is statistically significant different between the groups (also in case of	- "There are…no clear guidelines about the applied force. The 'European guidelines for quality assurance in breast cancer screening and diagnosis' …for example say that: 'the compression of the breast tissue should be firm but tolerable', but no concrete values are given." - In this study, only medio lateral oblique (MLO) images were used. In the screening program, MLO and cranio caudal (CC) images were acquired in the first screening round; in subsequent screening, CC images were acquired only in ≈57% pf patients until 2014 when it

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				>12.81 kPa: 4.9 (3.9-5.8) Pearson's χ^2: 0.011 • **PPV, % (95% CI)** ≤7.68 kPa: 25.4 (21.5–29.2) >7.68, ≤9.18 kPa: 31.2 (27.1–35.4) >9.18, ≤10.71 kPa: 32.7 (28.6–36.9) >10.71, ≤12.81 kPa: 25.8 (21.9–29.7) >12.81 kPa: 22.0 (18.4–25.6) Pearson's χ^2: 0.001	applying Bonferroni correction for multiple testing), it can be observed that a better performance is observed for most measures for the groups with a moderate pressure (group 2+3), compared to the other groups. These findings suggest, that a too low or too high compression may reduce screening program performance." **Author Reported Limitations** • "In this work we used a binning that created five groups, each containing 20% of the exams. The bin width is however not the same for all groups. Especially the first and the last group cover a large range of values. An alternative binning, based on the compression pressure distribution, will be investigated."	became obligatory to obtain two views in all rounds.
Other Quality Assurance Practices: Batch Reading						

<table>
<tr><th>1st Author, Date

[Country]</th><th>Reason for exclusion</th><th>Study/Participant Characteristics</th><th>Potential Influencing Factors of Recall Rate</th><th>Quantitative Results</th><th>Limitations and Conclusions</th><th>Comments</th></tr>
<tr>
<td><b>Burnside, 2005</b>
[USA]</td>
<td>• N/A</td>
<td>• Program: three fixed-site mammography facilities

• Study period: from October 2001 to July 2003.
First phase (October 1, 2001 to February 15, 2003): non-batch offline method of interpretation
Second phase (February 16, 2003 to July 30, 2003): batch reading offline

• Target age: not reported

• Screening frequency: not reported

• Sample size (# of screens) = 7,984 (before-batch reading); 1,538 (batch reading)

• Mean age (years): 56.2 (SD, 11.2) before batch reading; 56.6 (SD,</td>
<td>• Factor of Study: Batch Reading

• Other potential influencing factors:
<b>Reading Approach:</b> factor of study. ≈20-40 cases were evaluated per day using non-batch or batch approach.

<b>Readers’ Training:</b> five board-certified radiologists who fulfilled the MQSA requirements in terms of annual reading volume and continuing education and who interpreted at least 100 screening cases before and after the introduction of batch reading (2 fellow-trained in breast imaging; 2 practiced predominantly in other specialties; one general radiologist).</td>
<td>• <b>Recall Rate, % (95% CI)</b>

Radiologist
<table>
<tr><th>Radiologist</th><th>Before batch read</th><th>After</th></tr>
<tr><td>1</td><td>17.4 (16.0–18.9)</td><td>13.5 (10.8–16.5)</td></tr>
<tr><td>2</td><td>12.5 (11.0–14.1)</td><td>11.4 (8.3–15.1)</td></tr>
<tr><td>3</td><td>23.3 (20.1–26.7)</td><td>22.1 (16.5–28.5)</td></tr>
<tr><td>4</td><td>38.5 (35.2–42.0)</td><td>32.4 (24.6–40.8)</td></tr>
<tr><td>5</td><td>22.7 (20.8–24.8)</td><td>15.9 (11.5–21.0)</td></tr>
<tr><td>Total</td><td>20.1 (19.2–20.9)</td><td>16.2 (?)</td></tr>
<tr><td></td><td colspan="2">P<0.001</td></tr>
</table>
The decrease in recall rates was statistically significant for radiologists #1 and #5 who were fellowship trained in breast imaging.

Technology
<table>
<tr><th>Technology</th><th>Before batch read</th><th>After</th></tr>
</table>
</td>
<td><b>Author Reported Conclusions</b>
• “Our experience shows that batch reading can significantly reduce screening mammography recall rates without affecting the cancer detection rate or the proportion of cancers diagnosed with favorable prognostic indicators.”
<b>Author Reported Limitations</b>
• “…we did not randomize patients between the study groups. For this reason, we showed that the patient populations were not significantly different in terms of age, family history of breast cancer, and available comparison in order to confirm that they were unbiased.”
• “…although we have shown a definite improvement in recall rate for analog mammography, the effect of batch reading on digital</td>
<td>• “Dedicated batch reading requires an uninterrupted block of time designated to interpret a group of screening mammograms in succession”
• “Non-batch reading offline refers to interpreting screening mammograms in the midst of other duties such as diagnostic mammography or procedures after the patient has left the premises.”
• “Non-batch reading online entails interpreting mammograms with similar interruptions while the patient waits for her results.”
• “
• “There was no statistically significant difference in age,</td>
</tr>
</table>

<table>
<tr><th>1st Author, Date
[Country]</th><th>Reason for exclusion</th><th>Study/Participant Characteristics</th><th>Potential Influencing Factors of Recall Rate</th><th>Quantitative Results</th><th>Limitations and Conclusions</th><th>Comments</th></tr>
<tr><td></td><td></td><td>11.3) after adoption of batch reading</td><td>Radiologists who interpreted screening mammograms predominantly before or after the introduction of batch reading were excluded because comparison of their performance between the two sessions was not possible)

<b>Prior Mammograms:</b> not reported

• <b>Technology</b>: introduction of digital mammography in May 2002; computer assisted detection in October 2002 (analog) and in April 2003 (digital). Subset analyses were performed to control for these possible confounding variables.</td><td><table>
<tr><td>Analog</td><td>19.9</td><td>16.1</td></tr>
<tr><td></td><td colspan="2">P=0.009</td></tr>
<tr><td>Digital</td><td>21.0</td><td>16.3</td></tr>
<tr><td></td><td colspan="2">P=0.013</td></tr>
</table>
• <b>Cancer Detection Rate (per 1000)</b>
Radiologist
<table>
<tr><th>Radiologist</th><th>Before batch read</th><th>After</th></tr>
<tr><td>1</td><td>7.5 (5.0–11.0)</td><td>10.2 (4.0–22.0)</td></tr>
<tr><td>2</td><td>3.8 (2.0–8.0)</td><td>2.7 (0.0–15.0)</td></tr>
<tr><td>3</td><td>0 (0.0–6.0)</td><td>0 (0.0–18.0)</td></tr>
<tr><td>4</td><td>8.6 (3.0–18.0)</td><td>14.4 (0.1–51.0)</td></tr>
<tr><td>5</td><td>5.1 (2.0–10.0)</td><td>8.1 (0.1–29.0)</td></tr>
<tr><td>Total</td><td>5.6 (4.1–7.5)</td><td>7.2 (3.5–12.7)</td></tr>
<tr><td></td><td colspan="2">P=0.47</td></tr>
</table></td><td>mammography needs further clarification.”
• “CAD was added in October 2002 for analog images and April 2003 for digital images. During the 4-month period between introduction of analog CAD and the introduction of batch reading, a recall rate of 19.8% was recorded, slightly less than the prior 12 months of the study. Therefore, CAD did not inflate the recall rate before institution of batch reading. In addition, it is highly unlikely that CAD played a role in decreasing the recall rate during these or subsequent months because available evidence clearly establishes that recall rates are increased or unaffected by CAD…”</td><td>family history of breast cancer, or availability of prior mammograms between patients undergoing screening mammography before or after the institution of batch reading.”</td></tr>
</table>

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
Ghate, 2005 [USA]	• N/A	• Program: one of the four breast imaging facilities at the Department of Radiology, Duke University Medical Center • Study period: January 1 to October 31, 2001 • Target age: not reported • Screening frequency: yearly • Sample size (# of screens) = 8,698 • Mean age (years): 56.8±11.1 (immediate group); 56.2±11.3 (batch group). P=0.02	• Factor of Study: batch reading • Other potential influencing factors: **Reading Approach:** not reported **Readers' Training:** "five dedicated breast imaging radiologists" with 5 to 11 years of experience. The radiologists were rotated evenly between assignments for immediate and batch reading of the mammograms. **Prior Mammograms:** available for 83% of patients in the immediate group and 79% of patients in the batch group. P<0.001 • **Technology**: not clear; based on	• **Recall Rate (%)** Immediate group: 18% Batch group: 14% P<0.001 • **Cancer Detection Rate (%)** Immediate group: 0.49 (95% CI: 0.25, 0.73) Batch group: 0.43 (95% CI: 0.23, 0.63) P=0.7	**Author Reported Conclusions** • "In conclusion, immediate interpretation of screening mammograms results in higher recall rates, with no significant difference in cancer detection rates when compared with delayed subsequent batch interpreted mammograms." **Author Reported Limitations** • "An important limitation of our study is the possibility of selection bias. Because this is a retrospective database review, patients were not randomized between the two groups, and demographic characteristics between the two groups could not be directly controlled. … Our analysis determined that the two groups were closely matched with respect to breast	• "For immediate interpretations, images are evaluated and results are communicated with the patient at the time of the initial visit. Any necessary additional imaging is also performed during this visit." • "For radiologists…the unpredictable nature of immediate interpretation…can be disruptive to the flow of a busy diagnostic practice." • Batch reading: mammograms are read in a batch reading session after the patient leaves

1st Author, Date [Country]	Reason for exclusion	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			information from the "Discussion" section, screen-film images were digitalized and CAD was used.		density. The mean ages for the immediately interpreted group and the subsequent batch interpreted group were 56.8 years and 56.2 years, respectively. Although the mean ages were significantly different, the absolute difference of 6 months is very small, and it is likely of little clinical consequence." • "relatively small population…which reduces statistical power."	

Table A12. Radiologist characteristics

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
1st Author, Date *[Country]*	• *Program/Study Name* • *Study period* • *Target age* • *Screening frequency* • *Sample size* • *Age of women*	• *Factor of study: screen reader's characteristics* • *Other potential influencing factors: technology, quality assurance, reading approach, etc.*	• *Recall rate* • *False positive* • *Cancer detection rate* • *Positive predictive value*	*Conclusions* • *Author reported conclusions* *Limitations* • *Author reported limitations*	• *Comments (if any)*
Alberdi, 2011 [Spain]	• Program: four Spanish population-based breast cancer screening programs • Study period: March 1990 – December 2006 • Target age: 45-69 years • Screening frequency: biannual • Sample size (# of screens) = 1,440,384; (# women) = 471,112; (# radiologists) = 72 • Mean age (years): not reported	• Factor of Study: experience (years of service in the breast cancer screening program) • Other potential influencing factors: **Reading Approach:** Single reading **Reading Volume:** Only years in which radiologists interpreted at least 500 mammograms were included in these analyses **Technology**: "analog or digital, the latter being considered only if performed and read in a digital format"	• **Overall false positive (FP) OR (95% CI); multivariate analysis** <u>Years of service</u> <1 years (ref.) 1 year: 0.96 (0.93, 0.99); P=0.002 2 years: 0.86 (0.84, 0.89) P<0.001 3 years: 0.86 (0.83, 0.89) P<0.001 4 years: 0.79 (0.77, 0.82) P<0.001 >4 years: 0.72 (0.70, 0.74) P<0.001 • **FP leading to an invasive procedure OR (95% CI); multivariate analysis** <u>Years of service</u> <1 years (ref.) 1 year: 0.84 (0.77, 0.91) P<0.001 2 years: 0.62 (0.56, 0.69) P<0.001	**Author Reported Conclusions** • "…radiologists' length of service in the screening programme…reduced the risk of a false-positive result." • "this reduction was of a similar magnitude for overall false-positive results and for false-positives leading to an invasive procedure." • "…with overall false-positive results, the risk tended to decrease as the radiologist's length of service in the programme increased…" **Author Reported Limitations** • "…radiologist experience outside the screening programme was not taken into account."	• Article also reports on the effect of reading volume (see QA practices) • Data on radiologists' experience (years of service and reading volume) were obtained from screening program databases (in contrast to other studies that rely on self-reported data) • Cancer detection rates, PPV or sensitivity are not reported • Adjustment for the number of mammographic views, (one or two), mammogram type (analogue or digital), screen type (first or subsequent), period when the

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			3 years: 0.77 (0.69, 0.85) P<0.001 4 years: 0.75 (0.67, 0.84) P<0.001 >4 years: 0.73 (0.66, 0.80) P<0.001		mammogram was performed (in 2-year intervals), and patient's age. The radiology unit where the mammogram was performed was included in the model as a random effect.
Barlow, 2004 [USA]	• Program: mammography registries participating in the Breast Cancer Surveillance Consortium (BCSC) • Study period: January 1996 – December 2001 • Target age: ≥40 years • Screening frequency: not reported. To be included, a mammogram had to occur ≥9 months after any proceeding breast imaging to avoid misclassifying a diagnostic examination as screening • Sample size (# of screens) = 469,512; (# women) = 308,634; (# radiologists) = 124	• Factor of Study: age, gender, affiliation, experience, litigation concerns • Other potential influencing factors: **Reading Approach:** Not reported **Reading Volume:** Inclusion criterion: ≥480 mammograms annually over the study period **Patient Characteristic Considered:** Breast density, previous mammography, age, mammography registry **Technology:** not reported	• **Recall Rate (%)** Radiologist age (years): 35-44: 12.1 45-54: 10.6 ≥55: 8.5 Gender: Male: 9.8 Female: 11.4 Work full time No: 10.9 Yes: 9.9 Affiliation with an academic medical center Yes: 9.8 No: 10.3 Years of mammography interpretation <10: 11.8 10-19: 10.8 ≥20: 8.6 % of time spent working in breast imaging <20: 9.2 20-39: 11.3	**Author Reported Conclusions** • Radiologist's age, gender, malpractice experience, and malpractice concerns did not seem to be associated with performance." • "Radiologist's years of experience had the strongest association with performance, such that radiologists with fewer years in practice had higher sensitivity but lower specificity." **Author Reported Limitations** • "…the surveyed radiologists were not a random sample of all radiologists in the United States but only a sample participating in the national Breast Cancer Surveillance Consortium in three distinct locations."	• Article also reports data on the effect of reading volume (see QA practices) • For analysis by radiologists' demographics and experience, cancer detection rates or PPVs are not reported; therefore, data on sensitivity have been extracted • For analyses of litigation concern, recall rates or recall ORs are not reported; therefore, data on sensitivity and specificity have been extracted. • Final model: only sensitivity and specificity are reported

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
	• Mean age (years): not reported. Distribution by age categories (absolute numbers) is reported in table 1.		≥40: 9.5 • **Recall OR (95% CI); adjusted for patient's characteristics** Radiologist age (years): 35-44: 1.00 (ref.) 45-54: 0.88 (0.86, 0.9) ≥55: 0.78 (0.76, 0.8) P=0.001 Gender: Male: 1.00 (ref.) Female: 1.09 (0.86, 1.37) P=0.47 Work full time No: 1.00 (ref.) Yes: 0.95 (0.93, 1.04) P=0.11 Affiliation with an academic medical center Yes: 1.00 (ref.) No: 1.13 (0.83, 1.54) P=0.42 Years of mammography interpretation <10: 1.00 (ref.) 10-19: 0.81 (0.64, 1.02) ≥20: 0.66 (0.51, 0.85) P=0.005 % of time spent working in breast imaging <20: 1.00 (ref.) 20-39: 1.30 (1.06, 1.61) ≥40: 0.92 (0.68, 0.44) P=0.015		• Data on radiologists' experience are self reported

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			• **Sensitivity (95% CI); adjusted for patients' characteristics** Radiologist age (years): 35-44: 1.00 (ref.) 45-54: 0.79 (0.58, 1.10) ≥55: 0.52 (0.37, 0.74) P=0.001 Gender: Male: 1.00 (ref.) Female: 0.89 (0.66, 1.20) P=0.45 Work full time Yes: 1.00 (ref.) No: 0.60 (0.44, 0.82) P=0.002 Affiliation with an academic medical center Yes: 1.00 (ref.) No: 0.82 (0.52, 1.20) P=0.31 Years of mammography interpretation <10: 1.00 (ref.) 10-19: 0.69 (0.48, 0.98) ≥20: 0.50 (0.34, 0.74) P=0.003 % of time spent working in breast imaging <20: 1.00 (ref.) 20-39: 1.07 (0.78, 1.46) ≥40: 0.98 (0.65, 1.47) P=0.87		

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			• **Sensitivity (95% CI); adjusted for patients' characteristics** Medical malpractice insurance Self pay/other: 1.00 (ref.) Facility pays: 0.98 (0.60, 1.59) P=0.93 Ever had a malpractice claim No claims: 1.00 (ref.) Non-mammogram related: 0.79 (0.59, 1.06) Mammogram related: 0.86 (0.60, 1.22) P=0.28 Concerned about malpractice Disagree: 1.00 (ref.) Neutral: 0.77 (0.43, 1.35) Agree: 1.21 (0.73, 2.01) P=0.03 Malpractice influences recommendation for ultrasound Not changed: 1.00 (ref.) Moderately increased: 1.18 (0.87, 1.59) Greatly increased: 1.19 (0.77, 1.82) P=0.52		

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Interpreting mammograms is tedious Disagree: 1.00 (ref.) Neutral: 0.82 (0.57, 1.18) Agree: 0.94 (0.70, 1.28) P=0.56 Worry when not sure of a mammogram Disagree: 1.00 (ref.) Agree: 0.90 (0.67, 1.23) P=0.52 • **Specificity (95% CI); adjusted for patients' characteristics** Medical malpractice insurance Self pay/other: 1.00 (ref.) Facility pays: 0.96 (0.70, 1.32) P=0.81 Ever had a malpractice claim No claims: 1.00 (ref.) Non-mammogram related: 1.19 (0.97, 1.47) Mammogram related: 1.11 (0.84, 1.46) P=0.25 Concerned about malpractice Disagree: 1.00 (ref.) Neutral: 1.11 (0.73, 1.69) Agree: 0.87 (0.60, 1.25)		

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			P=0.14 Malpractice influences recommendation for ultrasound Not changed: 1.00 (ref.) Moderately increased: 0.92 (0.74, 1.15) Greatly increased: 0.81 (0.58, 1.13) P=0.46 Interpreting mammograms is tedious Disagree: 1.00 (ref.) Neutral: 0.95 (0.72, 1.25) Agree: 0.91 (0.73, 1.13) P=0.70 Worry when not sure of a mammogram Disagree: 1.00 (ref.) Agree: 0.88 (0.70, 1.10) P=0.25 • "Statistically significant radiologist factors were then tested together using mixed-effects models. Age of radiologist, percentage of time spent working in breast imaging, and concern about malpractice were no longer statistically significant after adjustment for other radiologist variables."		

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			- "The final model thus included the two radiologist factors of experience —i.e., number of years in mammography practice and number of mammograms interpreted per year." - **Sensitivity (95% CI); results from final model** Years of mammography interpretation <10: 1.00 (ref.) 10-19: 0.72 (0.52, 1.01) ≥20: 0.51 (0.36, 0.74) P=0.001 - **Specificity (95% CI); results from final model** Years of mammography interpretation <10: 1.00 (ref.) 10-19: 1.25 (1.00, 1.57) ≥20: 1.55 (1.21, 1.99) P=0.003		
Carney, 2004 [USA]	- Program: three mammography registries participating in the Breast Cancer Surveillance Consortium (BCSC) - Study period: 1 January 1996 – 31 December 2001	- Factor of Study: radiologist's reaction to uncertainty - Assessment of reaction to uncertainty - The reactions to uncertainty in the clinical decision-making instrument (adapted) included three scales: stress from	- Higher uncertainty scores were associated with increased recall rates, although the association was not statistically significant. - "…radiologists who are more experienced interpreters and those who interpret large volumes have lower	**Authors' Conclusion** - "Male radiologists report more intense reactions to uncertainty than do female radiologists, and reactions to uncertainty appear to lessen with more years of experience and with higher volume versus lower volume interpreters.	-

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
	• Screening frequency: not reported. • Sample size (# radiologists) = 124 • The radiologist survey was conducted to collect characteristics of radiologists (gender, years interpreting mammograms, reading volume, reimbursement mechanism, medico-legal experience, reaction to uncertainty in patient care) • Responses were linked to radiologists' performance data o from the three mammography registries.	uncertainty, concern about bad outcomes, and reluctance to disclose mistakes to physicians • Other potential influencing factors: gender, years interpreting mammograms, reading volume, reimbursement mechanism, medico-legal experience **Reading Volume:** Inclusion criterion: ≥480 mammograms annually over the study period **Technology**: not reported	reactions to uncertainty than do radiologists who are new to practice or who interpret smaller volumes of mammograms. A trend analysis suggests a trend for more years of interpretation and lower uncertainty scores, although this was not statistically significant (P = 0.13)." • "The mean combined uncertainty score was 33.5 (95% CI, 31.9 to 35.0) …It was lower among female radiologists (P = 0.01) as compared to male radiologists. More years interpreting mammography and higher interpretive volume were associated with lower uncertainty scores (P = 0.03)." • "Radiologists with any prior medico-legal experience had slightly higher uncertainty scores, although this was not significant (P = 0.31)."	Surprisingly, these 3 factors (gender, interpretive volume, and years in practice) were more closely associated with reactions to uncertainty than radiologists' medico-legal experiences." • "In conclusion, we found that radiologists interpreting screening mammography experience a range of reactions to uncertainty in their clinical practice. These reactions are higher than have been reported in other medical disciplines, such as internal medicine, and certain characteristics of the radiologist, such as gender and years of interpretive experience, mediate these reactions. Despite the high level of reactions experienced by radiologists to the uncertainty inherent in their practice, their interpretive performance appears to be unaffected." **Authors' reported limitations** • The study included radiologist practicing in three regions of the country. The findings may not be generalizable to	

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
				radiologists around the entire country • Inability to conduct analyses by specific practice settings, for example among radiologists whose practice is confined to mammography interpretation.	
Cornford, 2011 [UK]	• Program: East Midlands Breast Screening Programme • Study period: April 2005-March 2008 • Target age: not reported • Screening frequency: not reported • Sample size (# of screens) not reported; (# of film readers) = 37 (N=16 radiographers; N=21 radiologists)	• Factor of Study: years of film reading experience • Other potential influencing factors: **Reading Approach**: double-read with either consensus or arbitration for discordant cases; the second readers were not blinded to the results of the first reader **Readers' Training**: Radiographers with median film reading experience of 5.5 years (range 2-12 years) and median volume of film read during 3-year period of 13,163 (range 9864-19329) Radiologists with median film reading experience of 10 years (range 3-19	**Median (range)** • **Recall Rate (%)** <5 years: 6.2 (4.2, 9.8) 5 to >10 years: 5.9 (2.5, 10.4) 10 to <15 years: 6.7 (2.9, 8.7) 15 to <20 years: 3.6 (1.6, 7.4) • **Cancer Detection Rate per 1000** <5 years: 7.4 (5.8, 10.5) 5 to >10 years: 8.0 (5.4, 9.7) 10 to <15 years: 7.4 (6.7, 7.9) 15 to <20 years: 7.0 (5.6, 9.7) • **Small Cancer Detection Rate per 1000** <5 years: 4.1 (2.1, 5.8) 5 to >10 years: 4.2 (3.3, 5.1) 10 to <15 years: 4.2 (3.2, 5.0) 15 to <20 years: 4.0 (2.8, 5.2) • **PPV (%)** <5 years: 15.0 (5.9, 17.3) 5 to >10 years: 14.9 (8.8, 25.1) 10 to <15 years: 11.3 (9.1, 23.4) 15 to <20 years: 20.9 (11.9, 36.0)	**Author Reported Conclusions** • "The present study did not show an association between years of experience and any of the performance outcome measures." **Author Reported Limitations** • The results are likely to be affected by occupational group…" • Small sample size • "The present study was not large enough to examine the relationship between occupational group and either volume or experience"	• The units of recall rates are unclear: reported by the study authors as rates per 1000 but the values suggest these are rates per 100. • It appears that the analyses were not adjusted for patient's or reader's characteristics (likely due to small sample size)

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
		years) and median volume of film read during 3-year period of 22,538 (range 4,423-38,632) • All rates were calculated using first-reader data **Technology**: unclear (the terms "<u>film</u> readers" and "<u>film</u> reading" are used throughout the text)			
DiPrete, 2018 [USA]	• Program: one community practice in the USA • Study period: 2009-2014 Before experience with digital breast tomosynthesis (DBT): 2009-2011 After experience with DBT: 2012-2014 • Target age: not reported • Screening frequency: not reported • Sample size (# of screens) = 108,276 <u>Before DBT experience:</u> 50,062 <u>After DBT experience:</u>58,2014	• Factor of Study: the effect of the radiologist's experience with digital breast tomosynthesis on the radiologist's recall and cancer detection rate for routine two-dimensional digital mammography alone. • Other potential influencing factors: **Reading Approach:** Mammograms were batch-read by using CAD software (Cenova, version 1.3; Hologic; Bedford, Mass) **Readers' Training:** All six radiologists were fellowship-trained	• **Recall Rate (%)** <u>Before DBT experience</u>: Mean 6.8 Range 3.6-9.7 95% CI: 5.2, 9.0 Increase per year (2009-2011): 0.01% (P=0.9727) <u>After DBT experience</u>: Mean: 7.9 Range: 5.5-9.5 95% CI: 6.6, 9.3 Increase per year (2012-2014): 0.65%: (P=0.0127) <u>P (mean recall rate before vs. after DBT)</u> =0.0316 • **Detection Rate per 1000** <u>Before DBT experience</u>: Mean:2.5 95% CI: 2.2, 2.9 <u>After DBT experience:</u>	**Author Reported Conclusions** • "Recall rate, CDR, PPV2, and PPV3 of digital mammography increased after radiologist experience with DBT." **Author Reported Limitations** • "Although we know of no radiologist or institutional factor that changed during the study period that could affect recall rate or CDR, the retrospective design makes it impossible to control for the many factors that can alter recall rate." • "…because this was a single-institution study in which all six radiologists were fellowship trained in breast imaging, it may be difficult to generalize these results	• "Because of the increased cost and interpretation time associated with DBT… many imaging centers are gradually incorporating DBT into their clinical practices while still performing routine digital mammography alone in many patients." • In this study, the radiologists' performance was measured only in the community practice where only digital mammography was available, before and after DBT implementation at

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
	(# of radiologists) = 7 (one excluded due to low reading volume)	breast radiologists with 7-20 years of experience. All completed ≥8 hours of DBT in 2012; all worked both in a community practice in which only digital mammography was available and at two tertiary academic breast imaging centers where DBT was installed in 2012. **Technology**: digital mammography	Mean:3.5 95% CI: 3.0-4.0 P (CDR before vs. after DBT) = 0.203 • **PPV1 (%)** Before DBT experience: Mean:3.5 95% CI: 2.3, 5.4 After DBT experience: Mean:4.6 95% CI: 3.6, 5.9 P (PPV1 before vs. after DBT) = 0.1412	to a general radiology practice.” • The radiologists’ performance was investigated too soon after DBT introduction. • “Additional investigation of this topic after radiologists have gained more years of DBT experience is needed to further address the question of how DBT experience affects digital mammography interpretive performance.” • Small sample size	the academic practice. • Rates for each radiologist (n=7) before and after experience with DBT were calculated
Elmore, 2005 [USA]	• Data from three mammography registries contributing to the Breast Cancer Surveillance Consortium (BCSC) • Study period: January 1, 1996 to December 31, 2001 • Target age: not reported • Screening frequency: not reported • Sample size (# of screens) = 557 143; (# of women) = 308 634; (# of radiologists) = 124	• Factor of Study: litigation concerns • Other potential influencing factors: **Reading Approach**: not reported **Reading volume:** radiologists >80 screening mammograms in the BCSC database were considered for inclusion. **Readers’ Training:** radiologists **Prior Mammograms**: not reported **Technology:** not reported	• 64 radiologists (52.5%) reported a prior medical malpractice claim; 18 radiologists (14.8%) reported a mammography-related claims • Of 63 radiologists who reported a previous malpractice claim and responded to a question about the associated level of stress, 51 (81%) described the experience as very or extremely stressful. • 94 radiologists (76.4%) expressed concern about the impact medical malpractice has on their mammography practice.	**Author Reported Conclusions** • “U.S. radiologists are extremely concerned about medical malpractice and report that this concern affects their recall rates and biopsy recommendations. However, medical malpractice experience and concerns were not associated with recall or false-positive rates. Heightened concern of almost all radiologists may be a key reason that recall rates are higher in the United States than in other countries, but this	• “Univariate analyses were used to examine the associations between individual radiologist recall rates and medical malpractice perceptions and experiences. Mean recall rates and 95% confidence intervals were computed for each physician.” • “For each medical malpractice variable, the odds of recall were modeled by using logistic

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
	• Surveys were mailed to radiologists; response rate 76.8% • Mean age (years): not reported •	•	• 72 radiologists (58.5%) believed their concern moderately to greatly increased the number of recommendations for breast biopsies. • 43 radiologists (35.3%) considered withdrawing from mammography because of malpractice concerns. • Radiologists' estimates of malpractice risk were considerably higher than the actual risk. • **Malpractice experience and concerns and associated recall rates (95% CI) (%)** Ever have a medical malpractice claim No: 11.2 (10.2, 12.2) Yes, non-mammography related: 8.9 (7.7, 10.1) Yes, mammography-related: 10.1 (8.2, 12.0) If "yes", how stressful was it? Not at all or slightly: 5.8 (95% CI not calculated; N<5) Moderately: 11.7 (8.1, 15.4) Very: 9.4 (7.9, 10.8) Extremely: 8.6 (7.1, 10.1) Who pays for malpractice insurance?	hypothesis requires further study." **Author Reported Limitations** • "The surveyed radiologists were not a random sample of all U.S. radiologists" • "This study also did not include states with the highest medical malpractice activity." "data regarding malpractice were obtained by means of self-report, were for a short time period for some new radiologists, and were not verified; therefore, underreporting of malpractice experience was possible	regression while adjusting for study site."

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Facility: 10.2 (9.5, 11.0) Myself: 8.8 (5.7, 12.0) Other: not calculated (N=1) <u>I am concern about the impact of malpractice on my mammography practice</u> Strongly disagree: not calculated (N=0) Disagree: 12.3 (8.4, 16.2) Neutral: 9.1 (7.2, 11.0) Agree: 10.4 (9.3, 11.5) Strongly agree: 10.0 (8.8, 11.3) <u>How have malpractice concerns influenced your recommendations for diagnostic mammograms and/or US?</u> Greatly decreased: not calculated (N=0) Moderately decreased: not calculated (N=0) Not changed: 10.4 (9.1, 11.7) Moderately increased: 9.9 (8.9, 10.9) Greatly increased: 11.0 (8.9, 13.1) <u>How have malpractice concerns influenced the number of breast biopsies you recommended?</u>		

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Greatly decreased: not calculated (N=0) Moderately decreased: not calculated (N=0) Not changed: 10.3 (9.1, 11.5) Moderately increased: 10.1 (9.1, 11.1) Greatly increased: 10.1 (7.1, 13.1) How often do you consider withdrawing from mammography because of malpractice concerns? Not at all: 10.3 (9.3, 11.3) Yearly: 9.5 (8.0, 11.0) Monthly: 11.7 (9.0, 14.3) Weekly: 11.3 (8.7, 14.0) Daily: 8.1 (4.6, 11.6) How often do you consider withdrawing from general radiology because of malpractice concerns? Not at all: 10.0 (9.2, 10.9) Yearly: 10.2 (8.6, 11.8) Monthly: 11.4 (7.8, 15.0) Weekly: not calculated (N=1) Daily: 12.4 (95% CI not calculated; N<5) Perceived probability of malpractice suit in the next 5		

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			years if continue interpreting mammograms: **0**-19%: 10.2 (8.6, 11.8) 20-39%: 10.0 (8.3, 11.7) 40-59% 9.6 (8.4, 10.9) 60-79%: 11.3 (9.0, 13.6) 80-100%: 10.0 (8.2, 11.7) **Malpractice experience and concerns and associated odds of recall (95% CI)** Ever have a medical malpractice claim No: 1.00 (ref.) Yes, non-mammography related: 0.81 (0.68, 0.96) Yes, mammography-related: 0.90 (0.75, 1.07) If "yes", how stressful was it? Not at all or slightly: 1.00 (ref.) Moderately: 1.58 (0.65, 3.85) Very: 1.35 (0.58, 3.13) Extremely: 1.14 (0.49, 2.69) Who pays for malpractice insurance? Facility: 1.08 (0.77, 1.50) Myself: 0.91 (0.64, 1.28) Other: 0.48 (0.42, 0.55)		

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			<u>I am concern about the impact of malpractice on my mammography practice</u> Strongly disagree: -- Disagree: 1.36 (0.95, 1.96) Neutral: 1.00 (ref.) Agree: 1.22 (0.99, 1.50) Strongly agree: 1.12 (0.89, 1.39) <u>How have malpractice concerns influenced your recommendations for diagnostic mammograms and/or US?</u> Greatly decreased: -- Moderately decreased: -- Not changed: 1.00 (ref.) Moderately increased: 0.97 (0.81, 1.15) Greatly increased: 0.94 (0.73, 1.20) <u>How have malpractice concerns influenced the number of breast biopsies you recommended?</u> Greatly decreased: -- Moderately decreased: -- Not changed: 1.00 (ref.) Moderately increased: 0.97 (0.82, 1.14) Greatly increased: 0.87 (0.63, 1.20)		

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			How often do you consider withdrawing from mammography because of malpractice concerns? Not at all: 1.00 (ref.) Yearly: 0.92 (0.77, 1.10) Monthly: 1.13 (0.93, 1.39) Weekly: 1.02 (0.83, 1.27) Daily: 0.74 (0.50, 1.08) How often do you consider withdrawing from general radiology because of malpractice concerns? Not at all: 1.00 (ref.) Yearly: 0.96 (0.80, 1.14) Monthly: 1.07 (0.77, 1.48) Weekly: 0.57 (0.51, 0.65) Daily: 1.04 (0.94, 1.16) Perceived probability of malpractice suit in the next 5 years if continue interpreting mammograms: **0**-19%: 1.00 (ref.) 20-39%: 0.92 (0.72, 1.17) 40-59%: 0.87 (0.71, 1.08) 60-79%: 0.99 (0.78, 1.27) 80-100%: 0.86 (0.68, 1.09)		
Elmore, 2009 [USA]	• Program: seven Breast Cancer Surveillance Consortium (BCSC) sites	• Factor of Study: training, affiliation, experience, gender	Adjusted ORs (95% CI) • **Recall Rate** Gender Male: 1.00 (ref.) Female: 1.20 (1.00, 1.43)	**Author Reported Conclusions** • "Higher recall and false-positive rates were noted among female radiologists and radiologists with	• Adjustment for patients' characteristics (BCSC registry, age, breast density, time since

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
	- Study period: January 1, 1998 to December 31, 2005 - Target age: ≥40 years - Screening frequency: not reported - Sample size (# of screens) = 1,036,155; (# of women) = 531,705; (# of radiologists) = 205 - 257 of 364 eligible radiologists responded to a self-administered mail survey (71% response rate); survey results were linked to BCSC data on screening mammograms interpreted by these radiologists. Twenty-six radiologists with incomplete BCSC data were excluded.	- Other potential influencing factors: **Reading Approach:** not reported **Readers' Training:** radiologists (8% fellowship trained) **Technology**: not reported	P=0.047 Affiliation with academic medical center No: 1.00 (ref.) Yes, adjunct: 0.65 (0.48, 0.89) Yes, primary: 0.80 (0.61, 1.05) P=0.011 Fellowship training No: 1.00 (ref.) Yes: 1.45 (1.13, 1.86) P=0.004 Years of mammographic interpretation <10: 1.00 (ref.): 10-19: 0.90 (0.88, 0.93) ≥20: 1.05 (0.99, 1.11) P<0.001 Hours/week working in breast imaging 0-8: 1.00 (ref.) >8-16: 0.84 (0.69, 1.03) >16-32: 0.79 (0.61, 1.04) >32: 0.91 (0.73, 1.13) P=0.253 - **False positive rate** Gender Male: 1.00 (ref.) Female: 1.21 (1.01, 1.46) P=0.040	fellowship training, and lower recall and false-positive rates were noted among radiologists who had adjunct affiliations with an academic medical center and those with 10–19 years of experience interpreting mammograms… Higher sensitivity was noted for fellowship-trained radiologists, and PPV1 was lower among female radiologists." - "Fellowship training in breast imaging was the only characteristic of radiologists that was significantly associated with improved overall accuracy." - "In general, PPV1 was inversely associated with recall rate, but there was wide variability in PPV1, recall rate, and cancer detection rate across radiologists. Very few radiologists had both a high recall rate and a high PPV1. Radiologists with fellowship training tended to have higher PPV1 and cancer detection rates than did those without fellowship training." - "…we found that radiologists with fellowship	last mammographic examination), radiologists' random effect, radiologists' characteristics and reading volume - Unadjusted analyses: higher recall and false-positive rates were seen among female radiologists, radiologists with fellowship training in breast imaging, and those with <10 years of mammogram interpretation.

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Affiliation with academic medical center No: 1.00 (ref.) Yes, adjunct: 0.64 (0.46, 0.88) Yes, primary: 0.80 (0.61, 1.05) P=0.010 Fellowship training No: 1.00 (ref.) Yes: 1.45 (1.12, 1.87) P=0.005 Years of mammographic interpretation <10: 1.00 (ref.): 10-19: 0.90 (0.87, 0.93) ≥20: 1.06 (1.00, 1.13) P<0.001 Hours/week working in breast imaging 0-8: 1.00 (ref.) >8-16: 0.84 (0.68, 1.03) >16-32: 0.79 (0.60, 1.04) >32: 0.91 (0.72, 1.14) P=0.261 • **PPV1** Gender Male: 1.00 (ref.) Female: 0.81 (0.67, 0.97) P=0.025 Affiliation with academic medical center No: 1.00 (ref.)	training in breast imaging had significantly higher sensitivity and higher overall accuracy in screening mammograms than did non–fellowship-trained radiologists. However, these fellowship trained radiologists also had higher recall and false-positive rates." **Author Reported Limitations** • "low numbers of examinations in women with cancer… added to the variability we found in sensitivity • "…small number of fellowship-trained radiologists (n = 16) and the lack of data on the use of digital mammography." • "…30% of the study radiologists interpreted mammograms at institutions outside of the BCSC; thus, their self-reported data on annual volume could not be verified." • "…many of the radiologists worked part time, and this factor made interpretation of the percentage of time spent in breast imaging challenging.	

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Yes, adjunct: 1.25 (0.89, 1.77) Yes, primary: 1.05 (0.82, 1.36) P=0.416 <u>Fellowship training</u> No: 1.00 (ref.) Yes: 1.12 (0.83, 1.53) P=0.456 <u>Years of mammographic interpretation</u> <10: 1.00 (ref.): 10-19: 1.07 (0.96, 1.21) ≥20: 0.97 (0.81, 1.15) P=0.263 <u>Hours/week working in breast imaging</u> 0-8: 1.00 (ref.) >8-16: 1.24 (1.00, 1.52) >16-32: 1.10 (0.83, 1.45) >32: 1.09 (0.86, 1.38) P=0.218 • **Sensitivity** <u>Gender</u> Male: 1.00 (ref.) Female: 1.14 (0.83, 1.56) P=0.414 <u>Affiliation with academic medical center</u> No: 1.00 (ref.) Yes, adjunct: 0.60 (0.35, 1.04)		

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Yes, primary: 0.87 (0.57, 1.32) P=0.178 <u>Fellowship training</u> No: 1.00 (ref.) Yes: 2.32 (1.42, 3.80) P=<0.001 <u>Years of mammographic interpretation</u> <10: 1.00 (ref.): 10-19: 1.02 (0.81, 1.28) ≥20: 1.29 (0.95, 1.74) P=0.171 <u>Hours/week working in breast imaging</u> 0-8: 1.00 (ref.) >8-16: 0.74 (0.52, 1.05) >16-32: 0.69 (0.44, 1.09) >32: 0.67 (0.44, 1.01) P=0.228		
Miglioretti, 2009 [USA]	- Program: Seven mammography registries in the Breast Cancer Surveillance Consortium (BCSC) - Study period: January 1, 1996 to December 31, 2005 - Target age: 40-59 - Screening frequency: not reported	- Factor of Study: years of practice and fellowship training in breast imaging (the effect of fellowship training on the learning curve) - Other potential influencing factors: **Reading Approach:** 12% of the radiologists reported performing any double-reading; these	[Data on recall rates are reported in figures] <u>Radiologists without fellowship training</u> - **Recall Rate** "Mean recall and false-positive rates for radiologists with less than 1 year of experience were significantly higher than the AHRQ desirable goals…" "Mean recall and false-positive rates consistently	**Author Reported Conclusions** - "Radiologists with fellowship training in breast imaging experienced no learning curve and reached desirable [AHRQ] goals during their 1st year of practice." - "Radiologists without fellowship training in breast imaging significantly improved in their interpretation of screening mammograms as they	- "…digital mammograms were excluded because a separate learning curve could be associated with the introduction of this new technology." - "We examined the within-radiologist effect of increasing years of experience on the performance measures, separately

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
	- Sample size (# of screens) = 1,599,610; (# of women_ = 819,418; (# radiologists) = 231 from 280 facilities	radiologists reported double-reading for <5% of the mammograms. **Readers' Training:** radiologists with (7.4%) and without (92.6%) fellowship training in breast imaging **Prior Mammograms:** comparison films were available for 84.4% mammograms **Technology:** screen-film	met the AHRQ desirable goals only for radiologists with 19 or more years of experience." The largest improvement in the interpretive performance occurred during the first 3 years of practice: the odds of recalling a patient without cancer decreased by 11-15% per year. - **Sensitivity** "There were no significant trends in sensitivity with increasing years of experience." - **PPV** For radiologists with <1 year of experience, the mean PPV1 was significantly lower than the AHRQ desirable goal. "The average PPV1 fell within the AHRQ desirable range only for those radiologists with either 22 or 24 or more years of experience" <u>Radiologists with fellowship training</u> "Radiologists with fellowship training in breast imaging experienced no learning curve and reached desirable goals during their 1st year of practice." - **Recall Rate, PPV**	gained clinical experience following residency, while radiologists who received fellowship training in breast imaging did not have this learning curve in clinical practice. For radiologists without fellowship training in breast imaging, false-positive rates decreased sharply within the 1st 3 years of clinical practice, without evidence of an associated decrease in sensitivity." - "This learning curve… appears to continue well into a radiologist's career." - "Educational interventions, such as academic detailing and interactive case-based continuing education; system-level support, such as double reading with consensus and arbitration; and direct feedback on radiologists' interpretive performance through audit data may be especially important during the first few years of practice." **Author Reported Limitations** - The study was restricted to radiologists who agreed to respond to the mailed survey; however, the response rate was high and	for radiologists with and those without fellowship training in breast imaging, by using multivariable conditional logistic regression models that adjusted for patient age, mammographic breast density, time since last mammogram, whether comparison films were available, and whether CAD was used."

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			"There were no significant trends in recall rate (P=0.56), false-positive rates (P=0.58) or PPV1 (P=0.08) with increasing years of experience." - **Sensitivity** "Sensitivity decreased below 85%, though not significantly, for fellowship-trained radiologists with 2-3 years of practice but increased again for more experienced radiologists."	the interpretive performance of the participating radiologists was similar to the performance of the entire BCSC population. - Lack of adjustment for double reading; however double reading was rare. - "…we could not examine whether the learning curve depended on other factors, such as annual interpretive volume, where the radiologist trained, the type of feedback provided, or continuing medical education obtained."	
Smith-Bindman, 2005 [USA]	- Program: data from four mammography registries participating in the Breast Cancer Surveillance Consortium (BCSC) - Study period: January 1, 1995 to December 31, 2000 - Target age: not reported - Screening frequency: not reported	- Factor of Study: age, experience - Other potential influencing factors: **Reading volume:** Only physicians who read >=480 mammograms per year were included. **Reading Approach:** **Readers' Training:** "physicians" **Prior Mammograms: not reported**	- **False-positive rate** [reported in figures] "False-positive rate declined (i.e., specificity improved) with increasing physician age, with increasing time since receipt of medical degree, and with increasing annual volume. For example, among subsequent screening mammograms… the false-positive rate was 10.3% among physicians younger than age 40 years but only	**Author Reported Conclusions** - "In general, the most experienced physicians had the lowest false-positive rates. Physicians who had been practicing the longest, who interpreted 2500 – 4000 mammograms annually, and who emphasized screening, as opposed to diagnostic, mammography had lower false-positive rates than their less experienced counterparts. For physicians who had practiced the longest and who had a high focus on screening mammography, overall	- Adjustment for patient's and physician's characteristics (time since receipt of medical degree, average annual reading volume, ratio of screening to diagnostic mammograms) - This articles also reports on the effects of reading volume and ratio of screening to diagnostic mammograms

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
	- Sample size (# of screens) = 1,220,046' (# of physicians) = 209 - Mean age (years): not reported	**Technology:** not reported	6.8% among physicians aged 60-69 years.") - **Specificity (OR, 95% CI)** <u>Time since receipt of medical degree (years)</u> <10: 1.0 (ref.) 10-14: 1.16 (0.88 to 1.54). P=0.282 15-19: 1.22 (0.92 to 1.64). P=0.172 20-24: 1.18 (0.88 to 1.59). P=0.276 25-29: 1.54 (1.14 to 2.08). P=0.006 30-34: 1.67 (1.25 to 2.22). P<0.001 >34: 1.59 (1.12 to 2.22). P=0.008 - **Sensitivity (OR, 95% CI)** <u>Time since receipt of medical degree (years)</u> <10: 1.0 (ref.) 10-14: 0.98 (0.68 to 1.43). P=0.921 15-19: 1.07 (0.79 to 1.46). P=0.654 20-24: 0.96 (0.70 to 1.33). P=0.817 25-29: 1.00 (0.72 to 1.40). P=0.999 30-34: 0.86 (0.63 to 1.19). P=0.367	accuracy was improved as well, meaning that they had higher specificity without an equal loss in sensitivity." **Author Reported Limitations** - "…we do not know whether greater experience, higher annual volume, and a greater focus on screening mammography improve interpretations or whether the better physicians simply choose to interpret more examinations." - Sample size "was not large enough to look separately at ductal carcinoma in situ and invasive cancer".	

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			>34: 0.76 (0.55 to 1.04). P=0.084		
Tan, 2006 [USA]	• Data from Medicare claims (5% non-cancer sample) linked with the American Medical Association Masterfile (to obtain radiologists' characteristics) • Study period: 1998-1999 • Target age: >=65 years (Medicare beneficiaries) • Screening frequency: N/A • Sample size (# of screens) = 27,394; (# of women) = 21,576; (# of radiologists) = 1067 • Mean age (years): not reported; 60.57% of patients were 65-74-year old and 39.43% were older than 75 years	• Factor of Study: age, gender, experience • Other potential influencing factors: **Reading Approach:** **Readers' Training:** radiologists **Prior Mammograms:** **Technology** •	• **False Positive Rate (%)** <u>Age</u> <40 years: 7.55 40-49 years: 6.55 50-59 years: 5.57 60+ years: 5.27 P<0.001 <u>Gender</u> Female: 7.90 Male: 5.94 P<0.001 <u>Type of Practice</u> Direct patient care: 6.34 Indirect patient care: 6.26 P=0.89 <u>Years since graduation</u> <10: 7.92 10-19: 6.90 20-29: 5.84 30+: 5.27 P<0.001 <u>Board certification in radiology</u> Yes: 6.23 No: 6.62 P=0.46 • **False-Positive Rate: OR (95% CI); model included only radiologists' characteristics**	**Author Reported Conclusions** • "Radiologists varied greatly in accuracy of mammography reading. Female and more recently trained radiologists had higher false-positive rates. The variation among radiologists was largely due to unmeasured factors, especially unmeasured radiologist factors." **Author Reported Limitations** • Study limitations were primarily related to the use of claim data. • Screening mammograms may have been billed as diagnostic in claims data, and up to 20% of the screening mammograms may be missing from these analyses • Not all screening mammograms of Medicare beneficiaries are billed to Medicare. • "…radiologist's assessment and recommendations after a mammogram are not	• CDR, PPV or sensitivity are not reported

1st Author, Date [Country]	Study/Participant Characteristics	Potential Influencing Factors of Recall Rate	Quantitative Results	Limitations and Conclusions	Comments
			Gender Female: 1.25 (1.05, 1.49) Male: ref. Type of Practice Indirect patient care: 0.99 (0.80, 1.22) Direct patient care: ref. Years since graduation (per 10 years) 0.87 (0.81, 0.94) Board certification in radiology Yes: ref. No: 0.98 (0.78, 1.23) • **False-Positive Rate: OR (95% CI); model included patients' and radiologists' characteristics** Gender Female: 1.24 (1.04, 1.49) Male: ref. Type of Practice Indirect patient care: 0.99 (0.80, 1.22) Direct patient care: ref. Years since graduation (per 10 years) 0.87 (0.81, 0.94) Board certification in radiology Yes: ref. No: 0.99 (0.79, 1.24)	directly available in Medicare claims." • "measures of family history of breast cancer and radiologist volume of mammography could be underestimated."	

Appendix 6. Breast cancer screening programs: Organization and quality assurance practices.

Table A13. Breast cancer screening programs: Organization and quality assurance practices.

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
Centrally Organized Screening Program(s)	• Yes	• Yes	• Yes	• Yes • Organized screening program in France coexists with individual screening carried out at the initiative of the woman and her doctor (GP, gynecologist or radiologist)[14].	• No • "Rather than being invited, women can self-refer for screening and are advised to speak with their doctor to discuss screening appointments… Many insurance plans and providers remind their customers of the services that are available to them, and providers can market mammography directly to the public." (**Williams et al. 2015**)[15] • Medicare pays for mammography to women aged over 65 years.	• Yes • "In Canada, screening for breast cancer can occur within a cancer screening program (organized screening) or outside of such a program (opportunistic screening)." [16]

[14] Page 2 in Lastier D, Salines E, Danzon A. Programme de dépistage du cancer du sein en France : résultats 2007-2008, évolutions depuis 2004. Institut de veille sanitaire. Mai 2011. http://opac.invs.sante.fr/doc_num.php?explnum_id=7543

[15] Williams J, Garvican L, Tosteson AN, Goodman DC, Onega T. Breast cancer screening in England and the United States: a comparison of provision and utilisation. Int J Public Health. 2015 Dec;60(8):881-90. https://www.ncbi.nlm.nih.gov/pubmed/26446081

[16] Canadian Partnership Against Cancer. Breast Cancer Screening in Canada: Monitoring and Evaluation of Quality Indicators - Results Report, January 2011 to December 2012. Toronto: Canadian Partnership Against Cancer; 2017.

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
					• Medicaid and the National Breast and Cervical Cancer Early Detection Program pay for mammograms to low income women (**Williams et al. 2015**).	
Program Eligibility/Target Age	• Target age: 50-69 years (before 1 July 2013); 50-74 years (from 1 July 2013) • Eligible are women aged 40 years and above	• 50-69 years	• 50–70 years • "Some women outside this age group are also screened as part of the NHS Breast Screening Programme, either through self or General Practitioner (GP) referral where appropriate, or as part of a research trial."[17]	• 50-74 years	• **American College of Radiology**: 40+ (upper age is not indicated)[18] • **American Cancer Society**: from age 45 years for as long as overall health is good and the life expectancy is ≥10 years. Opportunity for screening should be given to women 40-44 years of age[19] • **U.S. Preventive Services Task Force**:	• 50-74 years

[17] Page 4 in Breast Screening Programme. England, 2015-16. https://digital.nhs.uk/catalogue/PUB23376

[18] Monticciolo DL, Newell MS, Hendrick RE, Helvie MA, Moy L, Monsees B, Kopans DB, Eby PR, Sickles EA. Breast Cancer Screening for Average-Risk Women: Recommendations From the ACR Commission on Breast Imaging. J Am Coll Radiol. 2017 Sep;14(9):1137-1143. https://www.ncbi.nlm.nih.gov/pubmed/28648873

[19] Oeffinger KC, Fontham ET, Etzioni R, Herzig A, Michaelson JS, Shih YC, Walter LC, Church TR, Flowers CR, LaMonte SJ, Wolf AM, DeSantis C, Lortet-Tieulent J, Andrews K, Manassaram-Baptiste D, Saslow D, Smith RA, Brawley OW, Wender R; American Cancer Society. Breast Cancer Screening forWomen at Average Risk 2015 Guideline Update From the American Cancer Society. JAMA. 2015;314(15):1599-1614. https://www.ncbi.nlm.nih.gov/pubmed/26501536

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
					50-74 years (women who place a higher value on the potential benefit than the potential harms may choose to start screening between 40 and 49 years)[20] • **The American Congress of Obstetricians and Gynecologists**: 40-75 years; the decision to discontinue after age 75 years should be based on a shared decision-making process informed by the woman's health status and longevity[21].	
Screening Intervals	• 2 years	• 2 years	• 3 years	• 2 years	• **American College of Radiology**: 1 year[22]	• 2 years for women aged 50+ years in all

[20] https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/breast-cancer-screening1

[21] https://www.acog.org/About-ACOG/News-Room/News-Releases/2017/ACOG-Revises-Breast-Cancer-Screening-Guidance--ObGyns-Promote-Shared-Decision-Making

[22] Monticciolo DL, Newell MS, Hendrick RE, Helvie MA, Moy L, Monsees B, Kopans DB, Eby PR, Sickles EA. Breast Cancer Screening for Average-Risk Women: Recommendations From the ACR Commission on Breast Imaging. J Am Coll Radiol. 2017 Sep;14(9):1137-1143. https://www.ncbi.nlm.nih.gov/pubmed/28648873

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
					- **American Cancer Society**: 1 year (45 to 54 years), 2 years (55+ years) [23] - **U.S. Preventive Services Task Force**: 2 years[24] - **The American Congress of Obstetricians and Gynecologists**: 1-2 years	provinces/territories except Nunavut[25] - 2 to 3 years recommended by the Canadian Task Force on Preventive Health Care.
Recall Rates PPV or Cancer Detection Rates [whichever reported] *Note: Age range for which recall rates are reported may be different from the*	Women aged 50-69 years. **Data for years 2003 to 2015** Recall rates - First screen: 9.4%, 9.8%, 9.8%, 9.9%, 9.9%, 9.9%, 10.7%, 11.1%, 10.7%, 10.8%, 11.6%, 12.2%, 11.7%	**Germany, North Rhine-Westphalia, women aged 50-69 years, year 2005-2009**[27] Recall rate - First screen: 6.1% - Subsequent screens: 3.4% Cancer detection rate (per 1000)	England, women aged ≥45 years **Year 2014-2015** Recall rates - Total: 4.2% - Prevalent Screens: 7.8% - Incident screens, last screen within 5 years: 3.0%	Women aged 50-74 years Recall rates[32] - 2004: 12% - 2005: 11.2% - 2006: 9.8% - 2007: 9.6% (first screens 12.3%, subsequent screens 8.3%) - 2008: 9.1% (first screens 12.2%;	**Abnormal Interpretations for 4,032,556 Screening Mammography Examinations from 1996- 2005 --- based on Breast Cancer Surveillance Consortium (BCSC) data**[35] Recall rate:	Women aged 50-69 years **Year 2003-2004**: Recall rate - Initial screen: 12.1% - Subsequent screens: 6.5% Positive predictive value (%) - Initial screen: 5.0%

[23] Oeffinger et al. Breast Cancer Screening forWomen at Average Risk 2015 Guideline Update From the American Cancer Society. JAMA. 2015;314(15):1599-1614. https://jamanetwork.com/journals/jama/fullarticle/2463262

[24] https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/breast-cancer-screening1

[25] See table 1 on page 7 in Canadian Partnership Against Cancer. Breast Cancer Screening in Canada: Monitoring and Evaluation of Quality Indicators - Results Report, January 2011 to December 2012. Toronto: Canadian Partnership Against Cancer; 2017

[27] Biesheuvel C, Weigel S, Heindel W. Mammography Screening: Evidence, History and Current Practice in Germany and Other European Countries. Breast Care (Basel). 2011 Apr; 6(2): 104–109. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3104900/

[32] Page 6 in Lastier D, Salines E, Danzon A. Programme de dépistage du cancer du sein en France : résultats 2007-2008, évolutions depuis 2004. Institut de veille sanitaire. Mai 2011. http://opac.invs.sante.fr/doc_num.php?explnum_id=7543.

[35] http://www.bcsc-research.org/statistics/benchmarks/screening/2007/table3.html

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
target age of the screening program	• Subsequent screens: 4.0%, 4.0%, 3.9%, 4.0%, 4.0%, 4.1%, 4.2%, 4.2%, 3.8%, 3.4%, 3.9%, 4.0%, 3.8% Invasive cancer detection rates (per 1000) • First screen: 6.35, 6.61, 6.45, 5.84, 6.36, 6.72, 6.52, 7.27, 7.07, 8.08, 8.31 • Subsequent screens: 4.44, 4.35, 4.24, 4.44, 4.30, 4.87, 4.68, 4.59, 4.44, 4.54, 4.94 Ductal carcinoma in situ (DCIS) detection rates (per 1000) • First screen: 1.58, 1.75, 1.41, 1.72, 1.87, 1.59, 1.8, 1.78, 1.86, 1.96, 2.41 • Subsequent screens: 1.04, 1.06,	• First screen: 3.1 • Subsequent screens: 2.3 **Italy, women aged 50-69 years, year 2007**[28] Recall rate • First screen 9.4% • Subsequent screens 4.1% Cancer detection rate (per 1000) • First screen 5.4 • Subsequent screens 4.8 **The Netherlands, women aged 50-75 years, year 2007**[29] Recall rate • First screen 3.5% • Subsequent screens 1.5%. Cancer detection rate (per 1000)	• Incident screens, last screen >5 years: 4.8% Cancer detection rate: • 8.6 per 1000 **Year 2015-2016** Recall rate: • Total: 4.1% • Prevalent Screens: 7.6% • Incident screens, last screen within 5 years: 3.0% • Incident screens, last screen >5 years: 4.6% Cancer detection rate: • 8.5 per 1000	subsequent screens 8.0%) • “Le taux du programme français peut donc apparaître élevé mais cela est lié aux modalités d'organisation spécifiques de ce programme: bilan de diagnostic immédiat, deuxième lecture des négatifs". [The rate of the French program may therefore appear high, but this is linked to the specific organization methods of this program: immediate diagnostic assessment,	• 10.9% Positive predictive value: • 4.4% **Abnormal Interpretations for 2,061,691 Screening Mammography Examinations from 2004 - 2008 -- based on BCSC data**[36] **through 2009** Recall rate: • 10.0% Positive predictive value: • 4.2% **Abnormal Interpretations for 1,682,504 Screening Mammography Examinations from 2007 – 2013 --** (Data from BCSC[37]) Recall rate:	• Subsequent screens: 7.3% Invasive cancer detection rate (per 1000) • Initial screen: 4.7 • Subsequent screen: 3.7 In situ cancer detection rate (per 1000) • Initial screen: 1.3 • Subsequent screen: 1.0 **Year 2005-2006** Recall rate • Initial screen: 12.2% • Subsequent screens: 6.0% Positive predictive value • Initial screen: 4.7% • Subsequent screens: 7.8%

[28] Data from tables 2 and 3 in Biesheuvel C, Weigel S, Heindel W. Mammography Screening: Evidence, History and Current Practice in Germany and Other European Countries. Breast Care (Basel). 2011 Apr; 6(2): 104–109. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3104900/

[29] Data from tables 2 and 3 in Biesheuvel C, Weigel S, Heindel W. Mammography Screening: Evidence, History and Current Practice in Germany and Other European Countries. Breast Care (Basel). 2011 Apr; 6(2): 104–109. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3104900/

[36] http://www.bcsc-research.org/statistics/benchmarks/screening/2009/table3.html

[37] http://www.bcsc-research.org/statistics/benchmarks/screening/2013/table3.html

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
	1.11, 0.98, 1.12, 1.17, 1.18, 1.17, 1.15, 1.13, 1.32 • “…the increase in the proportion of women who were recalled for further investigation in the last few years has led to an increase in the detection of breast cancer and DCIS for women screening for the first time. In this respect, the increase in the recall to assessment rate to above 10% for the first screening round may be considered acceptable.”[26]	• First screen 5.9 • Subsequent screens 5.2 **The Netherlands, women aged 49-74 years, year 2004-2010**[30] Recall rates • Digital mammography (DM): 2% • Screen-film mammography (SFM) in combination with DM at some point in time: 1.6% • SFM only: 1.6% Cancer detection rate (per 1000) • Digital mammography (DM): 5.9 • Screen-film mammography (SFM) in		second reading of the negatives.] [33] Cancer detection rates (per 1000)[34] • 2004: 7.8 • 2005: 7.4 • 2006: 6.8 • 2007: 6.7 • 2008: 6.3	• 11.6% Positive predictive value: • 4.4% **Cancers for 1,682,504 Screening Mammography Examinations from 2007 – 2013 --** (Data from BCSC[38]) Cancer detection rate: • 5.1 per 1000 • “…the BCSC is a collaborative network of seven mammography registries with linkages to tumor and/or pathology registries and supported by a Statistical	Invasive cancer detection rate (per 1000) • Initial screen: 3.8 • Subsequent screen: 3.2 In situ cancer detection rate (per 1000) • Initial screen: 0.5 • Subsequent screen: 0.6 **Year 2007-2008** Recall rate • Initial screen: 12.6% • Subsequent screens: 6% Positive predictive value • Initial screen: 4.8% • Subsequent screens: 7.7%

[26] BreastScreen Australia monitoring report 2012–2013

[30] van Luijt PA, Fracheboud J, Heijnsdijk EA, den Heeten GJ, de Koning HJ; National Evaluation Team for Breast Cancer Screening in Netherlands Study Group (NETB). Nation-wide data on screening performance during the transition to digital mammography: observations in 6 million screens. Eur J Cancer. 2013 Nov;49(16):3517-25. https://www.ncbi.nlm.nih.gov/pubmed/23871248

[33] Page 6 Lastier D, Salines E, Danzon A. Programme de dépistage du cancer du sein en France : résultats 2007-2008, évolutions depuis 2004. Institut de veille sanitaire. Mai 2011. http://opac.invs.sante.fr/doc_num.php?explnum_id=7543

[34] Table 4 in Lastier D, Salines E, Danzon A. Programme de dépistage du cancer du sein en France : résultats 2007-2008, évolutions depuis 2004. Institut de veille sanitaire. Mai 2011. http://opac.invs.sante.fr/doc_num.php?explnum_id=7543

[38] http://www.bcsc-research.org/statistics/benchmarks/screening/2013/table4.html

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
		combination with DM at some point in time: 5.1 • SFM only: 5.0 **Sweden, Stockholm county, women aged 40-69 years, year 1989-2008**[31] Recall rate • ≈3% Cancer detection rate • ≈0.5%			Coordinating Center” (**National Cancer Institute**[39]).	Invasive cancer detection rate (per 1000) • Initial screen: 4.7 • Subsequent screen: 3.7 In situ cancer detection rate (per 1000) • Initial screen: 1.2 • Subsequent screen: 0.9 **Year 2011-2012** Recall rate • Initial screen: 15.5% • Subsequent screens: 7.2% Positive predictive value • Initial screen: 4.1% • Subsequent screens: 6.5% Invasive cancer detection rate (per 1000) • Initial screen: 4.9 • Subsequent screen: 3.7

[31] Lind H, Svane G, Kemetli L, Törnberg S. Breast Cancer Screening Program in Stockholm County, Sweden - Aspects of Organization and Quality Assurance. Breast Care (Basel). 2010;5(5):353-357. https://www.ncbi.nlm.nih.gov/pubmed/21779220

[39] https://breastscreening.cancer.gov/

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
						In situ cancer detection rate (per 1000) • Initial screen: 1.2 • Subsequent screen: 0.8 **Year 2013-2014** Recall rate • Initial screen: 16.6% • Subsequent screens: 7.6% Positive predictive value • Not reported Invasive cancer detection rate (per 1000) • Not reported In situ cancer detection rate (per 1000) • Not reported [rates vary by province]
Target Performance Indicator for Recall Rate	• <10% (first screen) • <5% (subsequent screens)[40]	Acceptable level • <7% (initial screening)	Minimum standard • <10% (prevalent screen)	Acceptable rate • <7% (first screening)	Recall rate • 11.5% Positive predictive Value	Abnormal call rate • <10% (initial screen) • <5% (subsequent screens)[45]

[40] The service also monitors and reports the positive predictive value of a recall to assessment (See NAS 2015, measures 2.6.5 and 2.6.6 on page 123

[45] Page 17 in Canadian Partnership Against Cancer. Report from the Evaluation Indicators Working Group: Guidelines for Monitoring Breast Cancer Screening Program Performance (3rd Edition). Toronto: Canadian Partnership Against Cancer; February 2013.

Program Characteristics	**Country**					
	Australia	**Europe**	**UK (NHSBSP)**	**France**	**USA**	**Canada**
		• <5% (subsequent screening) Desirable level • <5% (initial screening) • <3% (subsequent screening)	• <7% (incident screen) Achievable standard • <7% (prevalent screen) • <5% (incident screen)	• <5% (subsequent screening) Desirable rate • <5% (first screening) • < 3% (subsequent screening)[41]	• 4.4%. [BCSC Benchmarks[42]. based on BCSC data through 2013[43]] Recall rate • <10% Positive Predictive Value • 5-10% [the Agency for Healthcare Research and Quality (AHRQ) desirable goals for screening mammography; as cited in Miglioretti et al. 2009[44], figures 1 and 2]	Positive predictive value • ≥5% (initial screen) • ≥6% (subsequent screens) [46]
Quality Assurance Guidelines	• National Accreditation Standards (NAS) of	• European guidelines for quality assurance in breast cancer screening and	• Quality Assurance Guidelines for Breast Cancer Screening	• The program is based on specifications published in the Official Journal of	• The Mammography Quality Standards Act (MQSA)	• CAR Practice Guidelines and Technical Standards

[41] Page 6 in Lastier D, Salines E, Danzon A. Programme de dépistage du cancer du sein en France : résultats 2007-2008, évolutions depuis 2004. Institut de veille sanitaire. Mai 2011. http://opac.invs.sante.fr/doc_num.php?explnum_id=7543

[42] National average rates are recommended by the Breast Cancer Surveillance Consortium (BCSC) as the target rates for providers: see https://www.advisory.com/research/imaging-performance-partnership/the-reading-room/2013/09/benchmarking-screening-and-diagnostic-mammography

[43] http://www.bcsc-research.org/statistics/benchmarks/screening/2013/table5.html

[44] Miglioretti DL, Gard CC, Carney PA, Onega TL, Buist DS, Sickles EA, Kerlikowske K, Rosenberg RD, Yankaskas BC, Geller BM, Elmore JG. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology. 2009 Dec;253(3):632-40. https://www.ncbi.nlm.nih.gov/pubmed/19789234

[46] Page 22 in Canadian Partnership Against Cancer. Report from the Evaluation Indicators Working Group: Guidelines for Monitoring Breast Cancer Screening Program Performance (3rd Edition). Toronto: Canadian Partnership Against Cancer; February 2013.

Program Characteristics	**Country**					
	Australia	**Europe**	**UK (NHSBSP)**	**France**	**USA**	**Canada**
	BreastScreen Australia[47]	diagnosis. Fourth edition. 2006 - "Local quality assurance manuals should be in use, which should be based upon European or national documents."	Radiology. Second edition. NHSBSP Publication No 59. March 2011 - The Royal College of Radiologists. Guidance on screening and symptomatic breast imaging, Third edition.	21 December 2006, which defines and regulates its procedures, monitoring and evaluation.[48]		for Breast Imaging and Intervention[49]. - Canadian Partnership Against Cancer. Quality Determinants of Breast Cancer Screening with Mammography in Canada. Toronto: Canadian Partnership Against Cancer; February 2013.
Accreditation	- "Accreditation is the independent review of a Service's and/or the State Coordination Unit's (SCU) performance…"[50] - "The National Quality Management Committee (NQMC) is the national body	- "A robust and reliable system of accreditation is required for screening and symptomatic units…" - "A European process of voluntary accreditation of Breast Units, based on the fulfilment of	- No information found on accreditation. Quality Assurance (QA) visits every 3 to 5 years (see below)	- No information found	- "Only facilities that are accredited by ABs [Accreditation Bodies], or undergoing accreditation by ABs, may receive certificates from FDA or an FDA-approved State Certifying Agency (CA) to legally perform mammography."	- Voluntary Mammography Accreditation Program (MAP)[54]

[47] The revised accreditation system was implemented on 1 January 2017 following a review that was finalized in 2015. http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/accreditation

[48] https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000460656&dateTexte=20180418

[49] https://car.ca/wp-content/uploads/Breast-Imaging-and-Intervention-2016.pdf

[50] National Accreditation Standards (2015): http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/CA8C934AA0B7BA64CA257EFA001C67D7/$File/BSA%20NAS%20Commentary%20April%202017.pdf

[54] https://car.ca/patient-care/map/

Program Characteristics	**Country**					
	Australia	**Europe**	**UK (NHSBSP)**	**France**	**USA**	**Canada**
	responsible for reviewing applications for accreditation and annual data reports from BreastScreen Services and SCUs and making a final decision about whether a Service and/or SCU is accredited." [51] • Accreditation survey every four years[52]	mandatory requirements should be established. To give uniformity a standard database should be made available."			• Four ABs approved by FDA through April 28, 2020: The American College of Radiology; State of Arkansas; State of Iowa; State of Texas • Facilities are accredited when they first open and every three years thereafter. • "On an annual basis, FDA assesses the performance of the ABs themselves." **(US FDA)** [53]	
Responsibility for quality assurance	• In each service, a Designated Radiologist is responsible for all aspects of quality	• "Regional and local organisations for QA should exist..." • "Each screening unit should have a Quality Assurance	• The screening quality assurance service (SQAS) in the NHSBSP[56] • The NHSBSP Radiology Quality	• Radiologists are committed to train, perform a quality control of the mammogram reading of their practice, and	• No information found	• No information found

[51] National Accreditation Standards (2015): http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/CA8C934AA0B7BA64CA257EFA001C67D7/$File/BSA%20NAS%20Commentary%20April%202017.pdf

[52] According to NAS 2015; NAS 2008 required site visits every two to four years.

[53] US FDA at: https://www.fda.gov/Radiation-EmittingProducts/MammographyQualityStandardsActandProgram/FacilityScorecard/ucm559308.htm

[56] NHS. Public Health England. Programme Specific Operating Model for Quality Assurance of Breast Screening Programmes. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/653748/BREAST_PSOM.pdf

Program Characteristics	Country					
	Australia	**Europe**	**UK (NHSBSP)**	**France**	**USA**	**Canada**
	assurance in screen reading[55].	Manager - one nominated person responsible for the overall quality of the programme." • "…the Quality Assurance Manager must ensure that all [quality assurance] results are collated for the programme and should act as a liaison between the local programme and the wider regional and national quality assurance organisations."	Assurance Coordinating Committee[57] • Regional quality assurance director[58] • "Quality Assurance Reference Centres (QARCs) should thoroughly investigate all screening services that have recall rates above the minimum standard." [59] • "It is the responsibility of all medical and non-medical practitioners providing radiology services to monitor their team's and their	transmit the mammograms interpretation sheets to the management structure, as well as the mammograms they deem normal for second reading[61].		

[55] The roles and responsibilities of the Designated Radiologist are outlined in Appendix C of NAS 2015

[57] NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011.

[58] NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011.

[59] NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011.

[61] Page 2 in Lastier D, Salines E, Danzon A. Programme de dépistage du cancer du sein en France : résultats 2007-2008, évolutions depuis 2004. Institut de veille sanitaire. Mai 2011. http://opac.invs.sante.fr/doc_num.php?explnum_id=7543

Program Characteristics	Country					
	Australia	**Europe**	**UK (NHSBSP)**	**France**	**USA**	**Canada**
			own performance…"[60]			
Readers' Qualifications	• Radiologists, breast physicians, general practitioners and radiographers[62]	• Radiologists[63]	• Radiologists, breast physicians, advanced practice radiographers and consultant radiographers[64]	• Licensed radiologists	• Physicians licensed to practice medicine in a State AND a) certified in radiology or diagnostic radiology OR b) have had at least 3 months of documented formal training in the interpretation of mammograms and in topics related to mammography AND other requirements to education and experience as listed on pages 9-11 in the US FDA guidance of 2001[65]	• Radiologists[66].

[60] NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011.

[62] Qualifications, training and experience required for radiologist and non-radiologist screen readers in the BreastScreen Australia Program are outlined in Appendix C of NAS 2015

[63] Interpretation of mammograms is not listed among responsibilities of radiographers in section 3.7 of the European guidelines for quality assurance in breast cancer screening and diagnosis, 4th edition. Professional requirements to radiologists are listed in section 4.7 of the European guidelines for quality assurance in breast cancer screening and diagnosis, 4th edition.

[64] NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011.

[65] https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm094441.pdf

[66] Requirements for qualification and training are listed on page 5 in CAR Practice Guidelines and Technical Standards for Breast Imaging and Intervention. https://car.ca/wp-content/uploads/Breast-Imaging-and-Intervention-2016.pdf and on page 40 in Canadian Partnership Against Cancer. Quality Determinants of Breast Cancer Screening with Mammography in Canada. Toronto: Canadian Partnership Against Cancer; February 2013.

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
Mammogram Reading/Interpretation and Recall Policies	- Independent ("blind") by two (or more) readers; at least one should be a radiologist - A single recommendation on whether to recall for assessment - Resolution of discordant opinions: a third reader (should be a radiologist with a high level of expertise in screen reading) or a consensus	**Centralized programs**: - Independent double reading is recommended - "for the first screening round and until the performance of the radiologists can be fully assessed". - Resolution of discordant opinions: consensus or a third screening radiologist **Non-centralized programs**: - double reading is mandatory; second reading should be performed by radiologists who read ≥5,000 mammograms/year	- "Double reading of mammograms by two film readers is recommended and should be considered mandatory in units that have moved entirely to digital mammography. Inexperienced readers should be paired with experienced readers and, ideally, readers with high recall rates should be paired with readers who have below-average recall rates and low cancer miss rates." [68] - "All services with prevalent and/or incident recall rates higher than	- Double reading. - If no abnormality is detected, the mammogram is read by a second radiologist. When an anomaly is detected, the first radiologist immediately carries out a diagnostic checkup	- "In the United States, single reading, increasingly with CAD, is the norm." (**Williams et al. 2015**[71])	- "Double reading is not the standard of care"[72] - "The breast screening program for Newfoundland and Labrador has a double read program that performs a second radiologist read on at least ten percent of all images as a component of the overall quality assurance of the program. The images are selected by the mammography technologist and/or nurse examiner based on suspicion; in addition, other random images are selected for double read to make up the ten percent. The highest reader approach [abnormal if either read indicates abnormal]

[68] NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011.

[71] Williams J, Garvican L, Tosteson AN, Goodman DC, Onega T. Breast cancer screening in England and the United States: a comparison of provision and utilisation. Int J Public Health. 2015 Dec;60(8):881-90. https://www.ncbi.nlm.nih.gov/pubmed/26446081

[72] Page 42 in Canadian Partnership Against Cancer. Quality Determinants of Breast Cancer Screening with Mammography in Canada. Toronto: Canadian Partnership Against Cancer; February 2013

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
		• "…cases recalled by one or both radiologists should be reviewed by an expert radiologist who can arbitrate."[67] • Giordano et al. (2012) summarized data on 26 European screening programs as of 2007. Twenty-one program implemented double reading.	the minimum standard must carry out arbitration of all prevalent and/or incident recalls as a matter of routine." [69] • Arbitration is undertaken by a third image reader or by a panel of image readers. Requirements to staff undertaking arbitration can be found on page 5 in Public Health England. NHS Breast Screening Programme Guidance on who can undertake arbitration. 2016[70]			is used to resolve discordance. Using this approach, the recall rate went from 6.6% to 7.2% and there was a 3.9% increase in the number of cancers detected through screening…Provincial breast screening programs in Nova Scotia also double read ten percent of all screening mammograms." [73]
Audit/Feedback	• Quarterly reporting of individual screen reader's	• "In order to maintain radiological performance	• Quality Assurance (QA) visits every 3 to 5 years (starting	• The quality of the French mammographic chain is checked	• National Mammography Database provides "comparative	• "Ongoing monitoring should include volumes, demographic

[67]. See table 1 of the publication. Giordano et al. Mammographic screening programmes in Europe: organization, coverage and participation. J Med Screen. 2012;19 Suppl 1:72-82. http://journals.sagepub.com/doi/pdf/10.1258/jms.2012.012085

[69] NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011.

[70] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/548405/Arbitration_guidance.pdf

[73] Page 42 in Canadian Partnership Against Cancer. Quality Determinants of Breast Cancer Screening with Mammography in Canada. Toronto: Canadian Partnership Against Cancer; February 2013

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
	performance. The quarterly Quality Assurance (QA) report includes the reader's recall to assessment rate[74]. • The QA report is provided to the reader, to the Designated Radiologist and the Clinical Director of the Service. • The Designated Radiologist discusses the reader's QA report and recommends action if required. • "…in 2011, BREAST (Breastscreen REader Assessment	standards, it is vital that the radiologist has direct access to key performance indicators…" • "Feedback of results at all stages is an important learning and quality enhancing process and mechanisms should be in place to achieve this." • "Each programme must review its own results in order to understand its own performance…"[76]	from April 2017)[77]. • "As a minimum all film readers should formally audit their film reading performance and compare their results with those of their peers. If their individual recall rates when acting as first reader are satisfactory, readers should review all the cases they did not recall where women were subsequently proven to have cancer. If their recall rates are	twice a year by approved organizations, according to the most recent recommendations of the National Agency for the Safety of Medicines and Health Products (ANSM). • The management structures annually transmit to the Institute of Public Health Surveillance [l'Institut de veille sanitaire, InVS], in a standardized format, the anonymized data needed for the	information for national and regional benchmarking. Participants receive semiannual feedback reports that include important benchmark data such as cancer detection rates, positive predictive value rates and recall rates." [82]	distribution of clients, standardized cancer detection rates and standardized abnormal call rates. …Where possible programs should provide trend data and relate this to provincial trends. Programs should have ranges for radiologist performance and include systems to support and educate radiologists considered to be performing outside the designated ranges."[83] • The results of monitoring and evaluation using the

[74] Page 58 in NAS 2015. It appears that the requirement to include recall rates in the QA report was introduced in NAS 2015. NAS 2008 requires inclusion of cancer detection rates.

[76] The European guidelines for quality assurance in breast cancer screening and diagnosis (4th edition) suggest computerized audit systems such as the European Screening Evaluation Database (SEED) at https://www.cpo.it/en/seed/. However, this database has not been found.

[77] NHSBSP guidelines on a QA visits can be found in Appendix 5 of the NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011. More information on QA visits can be found in Section 4 of the Public Health England (PHE) Programme Specific Operating Model for Quality Assurance of Breast Screening Programmes at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/653748/BREAST_PSOM.pdf

[82] American College of Radiology: https://www.acr.org/Practice-Management-Quality-Informatics/Registries/National-Mammography-Database

[83] Page 40 in Canadian Partnership Against Cancer. Quality Determinants of Breast Cancer Screening with Mammography in Canada. Toronto: Canadian Partnership Against Cancer; February 2013.

Program Characteristics	Country					
	Australia	**Europe**	**UK (NHSBSP)**	**France**	**USA**	**Canada**
	STrategy) which uses a novel web-based software developed by Ziltron (Ziltron, San Diego, CA, USA) and presents sets of clinically relevant screening mammograms, was introduced. The software monitors the readers' ability to determine which mammograms are positive for cancer and which cases are normal, and provides immediate feedback on sensitivity, specificity, and ROC [receiver operating characteristic] data along with image-		too high, readers should also review all their false positive recalls." [78] - Readers should "participate in PERFORMS (Personal Performance in Mammographic Screening) or a similar approved radiology performance QA scheme for mammography" [79]. - The PERFORMS test was implemented in the NHSBSP in 1991. Screen readers interpret a standard set of mammograms and receive immediate	evaluation of the program.[81]		CBCSD [Canadian Breast Cancer Screening Database] can be used to enhance the quality of organized breast cancer screening programs in Canada"[84].

[78] NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011.

[79] NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011. Information about the voluntary self-assessment scheme PERFPRMS can be found in: Scott HJ, Gale AG. Breast screening: PERFORMS identifies key mammographic training needs. Br J Radiol. 2006 Dec;79 Spec No 2:S127-33. https://www.ncbi.nlm.nih.gov/pubmed/17209118

[81] Page 2 in Lastier D, Salines E, Danzon A. Programme de dépistage du cancer du sein en France : résultats 2007-2008, évolutions depuis 2004. Institut de veille sanitaire. Mai 2011. http://opac.invs.sante.fr/doc_num.php?explnum_id=7543

[84] Canadian Partnership Against Cancer. Breast Cancer Screening in Canada: Monitoring and Evaluation of Quality Indicators - Results Report, January 2011 to December 2012. Toronto: Canadian Partnership Against Cancer; 2017

Program Characteristics	Country					
	Australia	**Europe**	**UK (NHSBSP)**	**France**	**USA**	**Canada**
	specific feedback."[75]		feedback on their performance[80].			
Required Minimum Reading Volume	• 2,000	• "5,000 screening cases per year in centralised programmes. This applies to the radiologist carrying out second reading in the non-centralised programmes."	• 5000 screening and/or symptomatic cases per year	• Radiologists known as "first readers" must perform at least 500 mammograms per year. Second reading radiologists must commit to reading at least 1,500 mammograms per year as part of this second reading activity.	• 960 mammographic examinations during a 24-month period[85]	• Interpret/second read a preferred minimum of 1,000 mammograms per year with a note that a minimum of 480 reads per year is still accepted" • "ideally, radiologists should read at least 2,000 mammograms per year" [86]
Other Quality Assurance Practices	• Comparison of the current screen with a previous screen (where available)	• "Previous mammograms should be displayed at the time of screen reading if ever possible."	• "Previous mammograms should be available to readers at the time of screen reading." [87]		• "Making comparisons with prior images significantly reduces false-positive findings…BSUs [Breast Screening Units] in England always have access to prior images. In	• "Original images from previous studies should be made available for consultation and second opinion where practical. Where prior images are digitized, the original images should be available

[75] Soh BP1, Lee W, Kench PL, Reed WM, McEntee MF, Poulos A, Brennan PC. Assessing reader performance in radiology, an imperfect science: lessons from breast screening. Clin Radiol. 2012 Jul;67(7):623-8. https://www.ncbi.nlm.nih.gov/pubmed/22486992

[80] Soh BP1, Lee W, Kench PL, Reed WM, McEntee MF, Poulos A, Brennan PC. Assessing reader performance in radiology, an imperfect science: lessons from breast screening. Clin Radiol. 2012 Jul;67(7):623-8. https://www.ncbi.nlm.nih.gov/pubmed/22486992

[85] Item 6 in US FDA 2001: https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm094441.pdf

[86] Pages 40-41 in Canadian Partnership Against Cancer. Quality Determinants of Breast Cancer Screening with Mammography in Canada. Toronto: Canadian Partnership Against Cancer; February 2013.

[87] NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011.

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
					the USA, a woman would need to return to the same provider in order for these comparisons to be made consistently." (**Williams et al. 2015**[88])	for review upon request."[89]
Mammography technology and imaging techniques	- "A routine screening examination would consist of four images for each client (2 cranio – caudal and 2 medio-lateral oblique" - Between 2009 and June 2013, all BreastScreen services had been digitalized.	- "Screening mammography using two views of each breast (medio lateral oblique plus craniocaudal) has been proven to be more effective than single oblique view screening, particularly in the woman's first round."	- "Two-view mammography (mediolateral oblique and craniocaudal projections of each breast) is required at each attendance."[90]	- Two-view mammography - In 2014, 95% of screenings were made using digital technology.	- Typically, two-view (mediolateral oblique and craniocaudal views) bilateral examinations (**Domingo et al. 2016)** - As of December 2009, 60% of accredited mammography facilities were using FFDM (**Domingo et al. 2016**) - "conventional digital mammography has essentially replaced	- "bilateral, two-view mammogram" [92] - "Two types of mammography are currently used for breast cancer screening in Canada: screen-film mammography (SFM) and digital mammography."[93]

[88] Williams J, Garvican L, Tosteson AN, Goodman DC, Onega T. Breast cancer screening in England and the United States: a comparison of provision and utilisation. Int J Public Health. 2015 Dec;60(8):881-90. https://www.ncbi.nlm.nih.gov/pubmed/26446081

[89] Canadian Association of Radiologists. Breast-Imaging-and-Intervention-2016: https://car.ca/wp-content/uploads/Breast-Imaging-and-Intervention-2016.pdf

[90] The Royal College of Radiologists. Guidance on screening and symptomatic breast imaging, Third edition. 2013. https://www.rcr.ac.uk/publication/guidance-screening-and-symptomatic-breast-imaging-third-edition

[92] Canadian Partnership Against Cancer. Breast Cancer Screening in Canada: Monitoring and Evaluation of Quality Indicators - Results Report, January 2011 to December 2012. Toronto: Canadian Partnership Against Cancer; 2017.

[93] Canadian Partnership Against Cancer. Breast Cancer Screening in Canada: Monitoring and Evaluation of Quality Indicators - Results Report, January 2011 to December 2012. Toronto: Canadian Partnership Against Cancer; 2017.

Program Characteristics	Country					
	Australia	Europe	UK (NHSBSP)	France	USA	Canada
					film mammography as the primary method for breast cancer screening in the United States."[91]	

Sources of information:

Australia. Rates: National Cancer Control Indicators[94]; BreastScreen Australia monitoring report 2012–2013[95]. Program characteristics and quality assurance practices: BreastScreen Australia National Accreditation Standards, October 2015[96]; BreastScreen Australia National Accreditation Standards, revised in April 2008[97]. Other data: BreastScreen Australia monitoring report 2012–2013[98]

Europe. Rates: see footnotes. Program characteristics and quality assurance practices: European guidelines for quality assurance in breast cancer screening and diagnosis[99]; Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition--summary document. Ann Oncol. 2008 Apr;19(4):614-22. [100]

United Kingdom. Rates: Tables F, I1 and I2 in Breast Screening Programme. England, 2015-16[101]. Program characteristics and quality assurance practices: NHS Cancer Screening Programmes. Quality Assurance Guidelines for Breast Cancer Screening Radiology, Second edition. NHSBSP Publication No 59. March 2011[102]. The Royal College of Radiologists. Guidance on screening and symptomatic breast imaging, Third edition. 2013[103]. Programme Specific Operating Model for Quality Assurance of Breast Screening Programmes. 2017[104]. NHS Breast Screening Programme Consolidated standards. 2017[105]. Public Health England. NHS Breast Screening Programme Guidance on who can undertake arbitration. 2016[106]

[91] Siu AL; U.S. Preventive Services Task Force. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med. 2016 Feb 16;164(4):279-96. https://www.ncbi.nlm.nih.gov/pubmed/26757170

[94] https://ncci.canceraustralia.gov.au/screening/abnormal-breast-screen-assessment/recall-assessment

[95] https://www.aihw.gov.au/reports/cancer-screening/breastscreen-australia-monitoring-2012-2013/contents/table-of-contents

[96] http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/bsa-nas-comm

[97] http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/2E9E74940BFB677BCA257D71007BF9F4/$File/standards.pdf

[98] https://www.aihw.gov.au/reports/cancer-screening/breastscreen-australia-monitoring-2012-2013/contents/table-of-contents

[99] http://www.euref.org/european-Guidelines

[100] https://academic.oup.com/annonc/article/19/4/614/217783

[101] https://digital.nhs.uk/catalogue/PUB23376

[102] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/470579/nhsbsp59_QA_radiology_uploaded_231015.pdf

[103] https://www.rcr.ac.uk/publication/guidance-screening-and-symptomatic-breast-imaging-third-edition

[104] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/653748/BREAST_PSOM.pdf

[105] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/612739/Breast_screening_consolidated_standards.pdf

[106] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/548405/Arbitration_guidance.pdf

France. Institut National du Cancer. Le programme de dépistage organisé [107]; Lastier D, Salines E, Danzon A. Programme de dépistage du cancer du sein en France : résultats 2007-2008, évolutions depuis 2004. Institut de veille sanitaire. Mai 2011.[108]

USA. Rates: Breast Cancer Surveillance Consortium**[109].** Program characteristics and quality assurance practices: Williams J, Garvican L, Tosteson AN, Goodman DC, Onega T. Breast cancer screening in England and the United States: a comparison of provision and utilisation. Int J Public Health. 2015 Dec;60(8):881-90[110]; Monticciolo DL, Newell MS, Hendrick RE, Helvie MA, Moy L, Monsees B, Kopans DB, Eby PR, Sickles EA. Breast Cancer Screening for Average-Risk Women: Recommendations From the ACR Commission on Breast Imaging. J Am Coll Radiol. 2017 Sep;14(9):1137-1143[111]; Oeffinger KC, Fontham ET, Etzioni R, Herzig A, Michaelson JS, Shih YC, Walter LC, Church TR, Flowers CR, LaMonte SJ, Wolf AM, DeSantis C, Lortet-Tieulent J, Andrews K, Manassaram-Baptiste D, Saslow D, Smith RA, Brawley OW, Wender R; American Cancer Society. Breast Cancer Screening for Women at Average Risk 2015 Guideline Update From the American Cancer Society. JAMA. 2015;314(15):1599-1614[112]; U.S. Preventive Services Task Force 2016[113]; The American Congress of Obstetricians and Gynecologists 2017[114]; American College of Radiology[115] ; US FDA[116]; US FDA 2001[117]; Breast Cancer Surveillance Consortium[118];

Canada. Rates: PHAC 2008. Organized Breast Cancer Screening Programs in Canada. Report on Program Performance in 2003 and 2004[119]; PHAC 2011. Organized Breast Cancer Screening Programs in Canada. Report on Program Performance in 2005 and 2006.[120]; Canadian Partnership Against Cancer. Organized Breast Cancer Screening Programs in Canada: Report on Program Performance in 2007 and 2008. Toronto: Canadian Partnership Against Cancer; February, 2013[121]; Canadian Partnership Against Cancer. Breast Cancer Screening in Canada: Monitoring and Evaluation of Quality Indicators - Results Report, January 2011 to December 2012. Toronto: Canadian Partnership Against Cancer; 2017[122]; Program characteristics and quality assurance practices: The Canadian Task Force on Preventive Health Care. Recommendations on screening for breast cancer in average-risk women aged 40–74 years. CMAJ November 22, 2011 183 (17) 1991-2001[123]; CAR Practice Guidelines and Technical Standards for Breast Imaging and Intervention[124]; Canadian

[107] http://www.e-cancer.fr/Professionnels-de-sante/Depistage-et-detection-precoce/Depistage-du-cancer-du-sein/Le-programme-de-depistage-organise
[108] http://opac.invs.sante.fr/doc_num.php?explnum_id=7543
[109] Breast Cancer Surveillance Consortium (BCSC) data:
http://www.bcsc-research.org/statistics/benchmarks/screening/2007/table3.html
http://www.bcsc-research.org/statistics/benchmarks/screening/2009/table3.html
http://www.bcsc-research.org/statistics/benchmarks/screening/
[110] https://www.ncbi.nlm.nih.gov/pubmed/26446081
[111] https://www.ncbi.nlm.nih.gov/pubmed/28648873
[112] https://www.ncbi.nlm.nih.gov/pubmed/26501536
[113] https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/breast-cancer-screening1
[114] https://www.acog.org/About-ACOG/News-Room/News-Releases/2017/ACOG-Revises-Breast-Cancer-Screening-Guidance--ObGyns-Promote-Shared-Decision-Making
[115] https://www.acr.org/Practice-Management-Quality-Informatics/Registries/National-Mammography-Database
[116] https://www.fda.gov/Radiation-EmittingProducts/MammographyQualityStandardsActandProgram/FacilityScorecard/ucm559308.htm
[117] https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm094441.pdf
[118] http://www.bcsc-research.org/statistics/benchmarks/screening/
[119] http://www.phac-aspc.gc.ca/publicat/2008/obcsp-podcs-03-04/pdf/obcsp-podcs-03-04-eng.pdf
[120] http://www.phac-aspc.gc.ca/cd-mc/publications/cancer/obcsp-podcs05/pdf/breast-cancer-report-eng.pdf
[121] http://www.getcheckedmanitoba.ca/files/b-rep-pro-perf-07-08.pdf
[122] https://content.cancerview.ca/download/cv/prevention_and_screening/screening_and_early_diagnosis/documents/breast_cancer_screening_canada_monitoring_evaluating_report_2011_12p?attachment=0
[123] http://www.cmaj.ca/content/cmaj/183/17/1991.full.pdf
[124] https://car.ca/wp-content/uploads/Breast-Imaging-and-Intervention-2016.pdf

Partnership Against Cancer. Quality Determinants of Breast Cancer Screening with Mammography in Canada. Toronto: Canadian Partnership Against Cancer; February 2013[125]; Canadian Partnership Against Cancer. Report from the Evaluation Indicators Working Group: Guidelines for Monitoring Breast Cancer Screening Program Performance (3rd Edition). Toronto: Canadian Partnership Against Cancer; February 2013[126].

[125] https://content.cancerview.ca/download/cv/prevention_and_screening/screening_and_early_diagnosis/documents/manmmographyincanadapdf?attachment=0
[126] https://content.cancerview.ca/download/cv/prevention_and_screening/screening_and_early_diagnosis/documents/guidelinemonitoringbreastpdf?attachment=0