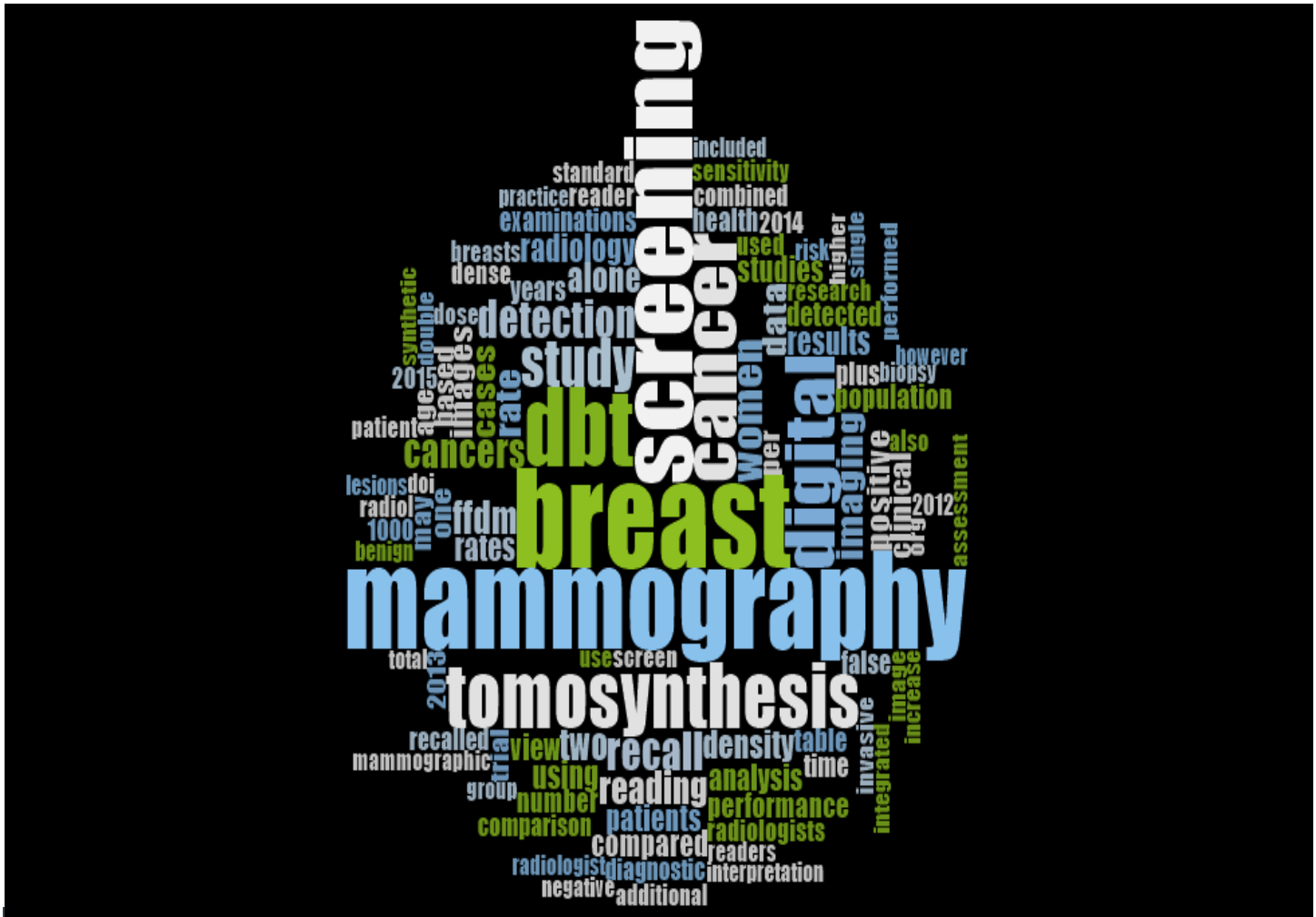


Digital breast tomosynthesis

A literature review to inform BreastScreen Australia's position statement on the use of tomosynthesis in screening

Final report: 30 April 2018

ALLEN+CLARKE

Document status:	Final report
Version and date:	V3, 30 May 2018
Author(s):	Anna Gribble, Stephanie James
Filing Location:	W/Department of Health/BreastScreen Australia/Position statement/Report
Peer / technical review:	Dr Robyn Haisman-Welsh
Verification that QA changes made:	Anna Gribble
Proof read:	Stephanie James
Formatting:	Anna Gribble
Final QA check and approved for release:	Anna Gribble

CONTENTS

KEY TERMS

95%CI	95% confidence interval
ASTOUND	Adjunct Screening with Tomosynthesis in Women with Mammography-negative Dense Breasts trial
BIRADS	Breast Imaging Reporting and Data System
BSA	BreastScreen Australia
CC	Craniocaudal (view)
CDR	Cancer detection rate
DBT	Two-view digital breast tomosynthesis (unless otherwise noted)
DBT_{CC}	One view digital breast tomosynthesis (craniocaudal view)
DBT_{MLO}	One view digital breast tomosynthesis (medio-lateral oblique view)
DM	Digital mammography
DM_{CC}	One view digital mammography (craniocaudal view)
FFDM	Full-field digital mammography (also known as two-view digital mammography)
JAFROC	Jackknife free-response receiver operating characteristic
MGD	Mean glandular dose
mGy	Milligray
MLO	Mediolateral oblique (view)
OR	Odds ratio
OTS	Oslo Tomosynthesis in Screening trial
PPV	Positive predictive value
PROSPR	Population-based Research to Optimize the Screening Process consortium
s2DM	Synthesised two-view digital mammography
STORM	Screening with Tomosynthesis or Regular Mammography trial

GUIDANCE ON HOW TO READ THIS REPORT

This report is a narrative literature review. It contains two main parts:

1. The *Key Findings* section provides a summary the findings of this literature review presented by the research questions. A summary of the evidence by clinical outcome and performance metric, and the GRADE assessment for key screening outcomes or metrics is also provided. We answer the research questions in a summary analysis in the Summary by research question section.
2. The main report provides detailed findings on to inform the research questions. Because many of the studies and articles included in this paper covered multiple screening outcomes and performance metrics (such as cancer detection rate, PPV, recall rate, etc.), we have presented the information by clinical outcome or performance metric rather than by study and article.

Appendix A includes the combined evidence tables for all the findings of this literature review. Appendix B includes the quality assessment tables (based on AMSTAR2 and the Scottish Intercollegiate Guidelines Network tools) for included articles.

KEY FINDINGS

The BreastScreen Australia (BSA) program currently uses bilateral full-field digital mammography (FFDM) as the "gold standard" screening test for the early detection of breast cancer in asymptomatic women aged over 40 years. The Department of Health (Australia) contracted *Allen + Clarke* to undertake a literature review (not systematic review) on the use of digital breast tomosynthesis (DBT) as a primary or adjunct screening test for the early detection of breast cancer in healthy, asymptomatic women. This review will support the Breast Screening Technical Reference Group's consideration of what updates (if any) are needed to BreastScreen Australia's position statement on DBT. Further updates to the BSA position statement may be required as recruiting or active studies report interim or final findings.

Most research discussed in this report used Hologic's DBT systems to acquire images. Where Hologic's systems were not used, Siemens Mammomat Inspirations system were used. Caution is needed if applying the findings reported in this literature review to other DBT systems (especially regarding radiation dose). Most studies reported on a dual acquisition strategy (that is, FFDM combined with DBT compared to FFDM). Other screening strategies included two-view DBT + synthesised 2D mammography (s2DM), or one-view DBT (mediolateral-oblique, DBT_{MLO}) compared to one-view digital mammography (craniocaudal – CC, DM_{CC}), FFDM or DBT_{MLO} + DM_{CC}.

Methodology

Allen + Clarke completed a systematic search of the Ovid Medline, CINAHL, Embase, ProQuest and Scopus databases as well as searches of health technology assessment, Cochrane and clinical trial databases. We used combinations of subject/index terms as appropriate to the search functionality of each database. Articles were included if they met the PICO(T) criteria. Studies focused on diagnostic populations were excluded from this review but will be explored in another literature review to investigate DBT's role in assessment and diagnosis.

The evidence base outlining the sensitivity, specificity, safety and acceptability of DBT as a primary or adjunct screening strategy is underpinned by data from five large prospective trials embedded in European population-based screening programs.[1] Other studies included data from the Population-based Research to Optimize the Screening Process (PROSPR) consortium, retrospective outcome analysis from single or multi-site community-based radiology practices, and retrospective observer performance studies. Information about the cost-effectiveness of DBT as a screening tool was not available but information about financial feasibility or other costs was reported in a small number of studies.

We found 85 relevant articles including two systematic reviews, 12 narrative literature reviews and 42 studies. Primary studies already incorporated into high-quality systematic or literature reviews were not assessed unless additional material not described in the systematic or literature review was included. An overall summary table provides an indication of the strength of findings.

[1] The Malmö trial (Sweden), the Screening with Tomosynthesis or Regular Mammography trials (STORM and STORM-2) (Italy), Adjunct Screening with Tomosynthesis or Ultrasound in Women with Mammography-negative Dense Breasts trial (ASTOUND) (Italy), and the Oslo Tomosynthesis in Screening (OTS) trial (Norway).

Results

Sensitivity

There is strong evidence that cancer detection rates (CDR) increase when using DBT compared to FFDM alone. Increases were reported in a range of studies (including large prospective trials) for different combinations of screening strategy including FFDM + DBT, DBT + s2DM, and DBT_{MLO} compared to DM_{CC} or FFDM alone. The direction of effect is consistent across study design, setting and location. There is variance in magnitude of effect.

CDR data from prospective trials

Pooled analysis based on data from the STORM and OTS trials reported a statistically significant increase of 2.43 cancers detected per 1000 screening examinations compared to FFDM alone. Adjunct screening with DBT also increases invasive cancer detection compared to FFDM alone. The incremental CDR and invasive CDR are similar for FFDM + DBT and DBT + s2DM: either detected significantly more cancers than FFDM alone. The Malmö trial also reported a significant increase in CDR: 2.6 cancers detected per 1000 screening examinations using DBT_{MLO} compared to FFDM alone.

DBT + s2DM also performed better, detecting 8.8 cancers per 1000 screening examinations compared to 6.3 with FFDM alone. The evidence of comparative CDR performance for DBT + s2DM compared to FFDM + DBT was inconsistent but the CDR is similar in both screening strategies (in both smaller and larger studies). DBT + s2DM is a promising screening strategy, especially as it significantly reduces radiation dose.

CDR data from retrospective analysis or reader studies

Data from retrospective studies showed a similar effect (that is, increases in CDR when DBT is used) but the increases were smaller than those reported from the prospective trials. Statistically significant CDR results from retrospective studies ranged from 1.6 to 1.9 cancers detected per 1000 screening examinations. Reasons for the lower CDR in the retrospective studies could relate to differences in reading strategy (eg, double reading compared to a single reader approach), participant selection, under-powering or other study design limitations.

Types of cancer detected

Earlier studies showed that FFDM + DBT's performance did not appear to be superior for the detection of ductal carcinoma in situ because of reduced visibility of microcalcifications. Later studies are reporting no differences in the types of cancers detected by either FFDM + DBT or FFDM alone. Further research is needed to determine DBT's ability to detect microcalcifications.

There was very limited data about the long-term mortality benefits, treatment morbidity or quality of life improvements associated with FFDM + DBT as a screening strategy. Almost no data exists on results for incident screening compared to prevalent screening, mortality benefit or surrogate indictors of this. Reliable data on interval cancer rate is also scarce.

Specificity

The literature is not settled about the association between DBT and recall rates.

Data from prospective trials

Some results show that overall recall rates can be reduced when using FFDM + DBT compared with FFDM alone. Other prospective trials reported increased recall with double reading (either

by two radiologists or through an arbitration process) but reduced false positive rate. This may reflect that the overall false positive recall rates within screening programs where prospective trials are embedded are generally low anyway.

There is less literature exploring associations between DBT + s2DM and the rate of false positive recalls although research generally favours a reduction in false positives with DBT + s2DM compared with both FFDM + DBT and FFDM alone. Like overall recall rates, there is some variance in the direction of effect. Results from a large prospective trial (STORM 2) showed a false positive recall rate for DBT + s2DM that was significantly greater than for FFDM + DBT and FFDM alone. It is possible that the results from the STORM-2 trial relate to early experiences of incorporating s2DM into real-world screening practice for the first time without previous experience with s2DM images relative to FFDM. Secondary analysis from the STORM 2 trial indicated that false positive recall rates for FFDM + DBT and DBT + s2DM significantly reduced compared to those for FFDM. These results reflect developing and increasing knowledge in the use of FFDM + DBT, with some interpretation issues still present for s2DM. Interim results from the Malmö trial reported a reduction in the false positive recall rate of screens using DBT over the first 1.5 years, which also indicates that false positive recall could be associated with a learning curve in interpretation.

Data from retrospective analysis or reader studies

Information from the smaller retrospective studies (most of which used single reading strategies) reported that recall rate was reduced with the addition of DBT to FFDM. Other trial data reported both reduced recall rate and reduced false positives with the addition or use of DBT. Differences in overall program false positive recall rates, reading strategy and arbitration protocols used to determine which women to recall from screening may account for some of the inconsistency. Increasing reader experience, knowledge of DBT and interpreting 3D images and availability of prior DBT images may also further decrease recall rates.

PPV

Overall results on PPV_1 indicated that FFDM + DBT accurately detected proportionally more women recalled from screening who had breast cancer compared to FFDM alone. DBT + s2DM also showed promise of increased accuracy. Screening based on DBT + s2DM screening may correctly identify between one and three more women with diagnosable breast cancer for every 100 women recalled, compared with recalls based on FFDM + DBT screening. Results on the PPV for biopsy recommended and biopsy performed indicate that FFDM + DBT was also more accurate than FFDM alone when used as a basis for recommending or performing biopsies. PPV_{2-3} results for DBT + s2DM are also promising but present more varied effect size than results for FFDM + DBT.

Safety

Radiation dose varies with the image acquisition process used (DBT or FFDM or combination mode), the number of and type of views, the use of automatic exposure control, positioning, breast size and composition, and by DBT system used.

Much of the published evidence about the sensitivity and specificity of DBT is based on dual acquisition protocols (i.e., FFDM + DBT). Using average breast thickness, the radiation dose required to acquire acceptable images with FFDM + DBT is approximately double that of FFDM alone (2.98mGy compared to 1.49mGy). This 'double dose' is still within the dose limits set for overseas quality and safety standards but is higher than the per view dose limit set for the BSA

program. The radiation dose for DBT compared with FFDM is lower: DBT_{MLO} has about 70% of the mean glandular dose (MGD) compared to FFDM alone. Two-view DBT results in a similar MGD compared with FFDM. Other possible single view combinations also result in lower MGD.

Synthesised acquisition of 2D images halves the effective dose of combined FFDM + DBT, making it comparable to FFDM alone but with the improved detection rates associated with DBT. Initial studies indicate that the quality of images reconstructed from s2DM is acceptable, but further evidence is required to ensure that they can be used to accurately interpret microcalcifications.

Having FFDM + DBT as the preferred screening strategy will have implications for cumulative dose if separate acquisitions are used for 2D and 3D images, if the screening interval is annual not biennial, or if women participate in mammography-based screening from their early 40s.

The effect of age, breast density and screening interval on sensitivity, specificity and safety

Findings for CDR stratified by breast density present results that may be surprising, given that DBT improves conspicuity and should, in theory, provide more quality images of more dense breasts. While results for FFDM + DBT show increased CDR for all women, data from prospective trials does not demonstrate a significant increase in CDR when comparing women with more dense breasts to those with less dense breasts. The use of breast density classification systems can result in unreliable allocation because density classification can be affected by factors like hormone levels, genetic factors, parity, use of oestrogen, place in menstrual cycle, use of tamoxifen, weight and (importantly) inter/intra reader variability. It is possible for women to be classified as having non-dense breasts (BIRADS 2) in one mammogram but be reclassified to having more dense breasts in the next mammogram (and vice versa). Research may therefore be comparing the most dense breasts to those that could be determined to be more fatty but still have significant areas of mammographic density. This could account for the smaller-than-expected increase in CDR between women with more dense or less dense breasts. Research which reports CDR, recall rates and false positives by “extremely dense” (BIRADS 4) and “almost entirely fatty” (BIRADS 1) could result in clearer (and possibly truer) results on CDR differences.

The age at which screening participation begins and screening interval may also influence sensitivity, specificity and overall lifetime radiation dose received. Studies reporting age stratification used different age bands (i.e., 10-year bands or groups like over 60 years/under 60 years). This impacts on our ability to draw useful conclusions about the possible relationships between age, and clinical outcomes or performance metrics associated with the use of DBT.

Reader performance and interpretation time

Most studies involved readers with a range of experience in breast screening and radiology in general. Less experienced readers improved more in interpretation accuracy compared to more experienced radiologists. It is unclear whether this improvement reflects the development of less experienced practitioners’ competence, or whether DBT is ‘easier’ to read without as much experience in breast cancer imaging.

DBT produces many more images than FFDM alone. Strong and consistent results indicated that implementing FFDM + DBT (or DBT alone) increased reading and interpretation times.

Increased reading and interpretation time has workflow, radiologist/reader resourcing and cost implications if DBT is implemented into population screening programs.

Financial implications

The literature is dominated by studies that used modelled analyses to discuss the effect of DBT implementation on insurance programs in the United States. Modelled analyses showed that FFDM + DBT demonstrated economic favourability when considering clinical benefits like cancer detection and recall rates. Unfortunately, due to differences in health sector policy and service delivery, the modelled analyses may have limited applicability to the BSA program. No modelled analyses focused on the implementation of DBT have been conducted elsewhere.

Implementation of DBT would require capital upgrade costs (eg, new equipment requirements), increased capacity for data storage or transmission, training and additional time for radiologists to read DBT images. Incremental costs could be offset by health system savings associated with increased cancer detection and reduced rates of recall (in particular, the costs associated with recalling women unnecessarily, additional unnecessary biopsies and further assessment in cases where breast cancer is not present). At the time of this review, no detailed cost analysis had been reported. Cost is still something that needs to be balanced against potential benefits.

Acceptability

There is little evidence describing the acceptability to women or practitioners of DBT as a screening tool. We infer that women may appreciate the lower compression that can be used to acquire acceptable images with DBT_{MLO} compared to FFDM, but women may be concerned about the radiation dose associated with dual acquisition. Images acquired by DBT using lower compression provided acceptable quality images for some, but not all, radiologists. Women's anxiety at participating in a screening program or receiving screening results may be reduced if CDR is improved (with DBT) and if false positive results are reduced and unnecessary recalls are avoided. Further research investigating DBT's acceptability to women and practitioners is required.

Are other jurisdictions implementing DBT into population-based screening programs?

Eight jurisdictions (including Europe) have current position papers on the role of DBT in screening. Most report the same conclusion: existing evidence favours FFDM + DBT compared to FFDM alone for key clinical outcomes like CDR and recall rates. It is a promising technology that will have some role in the future of screening programs. Concerns remain around the increased radiation dose associated with dual acquisition (which could be addressed by s2DM). The main issue is the lack of evidence about DBT's impact on long-term clinical outcomes and cancer mortality reduction. Jurisdictions generally recommend that further evidence from prospective trials and RCTs be used to inform decisions about DBT's integration into national screening programs (except for France, which is moving away from national population-based screening programs in favour of personalised screening based on improved informed consent).

A developing evidence base

A complex evidence base covering different combinations of screening approach underpins DBT as a promising imaging technology for the early detection of cancer in asymptomatic women. Overall, DBT (alone or as an adjunct screen to FFDM) performs better than FFDM alone. There is more limited, but promising, evidence that DBT + s2DM has a superior performance to FFDM alone, often achieving as-good results compared to FFDM + DBT but with a lower radiation

dose. The algorithms that reconstruct the 2D images from DBT data are improving as is reader performance and further improvements (including those related to access to prior DBT images with incident screening) could be achieved with s2DM in the future. Other ways of using DBT are being explored but have reported limited results to date. Future research could usefully focus on determining which combination of screening strategies involving DBT result in the maximum benefit to women (and to which women).

There is limited evidence of the longer-term impact of DBT on cancer mortality (including overall cost-effectiveness), treatment morbidity or quality of life improvement. Most studies have had follow-up of less than 24 months and long-term adequately powered studies are limited. Some data on proxy measures (such as tumour size at detection) is available but this is insufficient to provide a sense of the long-term mortality reduction benefits that DBT may offer compared to FFDM. Seven large active or recruiting trials embedded in population-based screening programs in Europe and Canada, as well as Australian research from the Maroondah trial, will provide further evidence on the efficacy, effectiveness and safety of DBT as a screening test (either alone, with s2DM or as an adjunct to FFDM). Information on financial cost would also be useful. The following quote from Maroondah trial information summarises the current evidence base:

> *"From a clinical point of view, tomosynthesis has demonstrated usefulness in cancer detection and characterisation and in reducing recall rates, but BreastScreen Australia, quite rightly, will not make a move until the final outcomes of larger screening trials on the use of tomosynthesis in a screening environment are known. We must watch and wait,"* Dr Lockie, Maroondah trial[2].

Summary by research question

Question 1: Should DBT (with or without s2DM) be used as the primary screening tool for the early detection of breast cancer in asymptomatic women aged over 40 years?

While DBT improves breast cancer detection with (potentially) lower rates of recall than FFDM alone, there is insufficient evidence about the long-term mortality benefit to support the use of DBT alone as a primary screening test. Few studies identified for this literature review investigated DBT alone compared to FFDM alone, although literature exploring different ways to integrate DBT into screening continues to develop quickly. There is promising data that DBT + s2DM performs comparably to that of FFDM + DBT but with a substantially reduced radiation dose. Another active trial is investigating the comparative performance of DBT_{MLO} in relation to FFDM. Additional studies on DBT alone (with or without s2DM) are required to more deeply explore performance of one-view DBT compared to that of DBT + s2DM or FFDM. Most of the upcoming research, however, does not focus on this screening strategy, possibly because of concerns that DBT alone may not detect some microcalcifications indicative of DCIS.

[2] http://www.breasttomo.com.au/sites/breasttomo.com.au/files/MAROONDAH%20BREASTSCREEN.pdf.

Question 2: For the early detection of breast cancer in asymptomatic women aged over 40 years, is FFDM + FFDM + DBT (including s2DM) a more efficacious and safer screening strategies than FFDM alone?

There is strong evidence that both FFDM + DBT and DBT + s2DM provide superior performance for improved cancer detection and that DBT may be more sensitive test than FFDM alone. The magnitude of improvement may be affected by reading strategy (eg, double reading approaches result in higher detection rates compared to single reading). The radiation dose per FFDM + DBT view is almost double that of FFDM alone (but is still below international dose limits).

There is emerging evidence that, used as an adjunct screen to FFDM, DBT can reduce recall rates and false positives results compared to FFDM alone; however, some inconsistent results between large prospective trials are reported (which could reflect the already low rates of recall seen in some population-based screening programs in which the trials are embedded). Further research investigating the comparative performance will help to unpick areas of uncertainty including the impact of double/single reading strategies and the impact of access to previous DBT images.

Further research may also help determine which combination of approaches (FFDM + DBT, two-view DBT + s2DM, DBT_{MLO} + FFDM, or some form of DBT alone) achieves the best balance between radiation dose, sensitivity and specificity. Further research is needed to confirm the association between screening strategies using DBT and outcomes like interval cancer rate and long-term mortality benefit, treatment morbidity or quality of life improvements. Finally, over-diagnosis may also be an issue requiring improvement in treatment decision tools rather than reduced initial detection (eg, differentiating which abnormalities may never be clinically significant and following appropriate management protocols rather than making it an issue of detection alone).

Question 3: For the early detection of breast cancer, are there population groups for which FFDM + DBT (including s2DM) is a more efficacious, safer screening strategy than FFDM alone?

There is limited evidence that FFDM + DBT could result in differential performance for different population sub-groups. There is insufficient evidence to determine the nature of DBT's performance in relation to women who have more dense breasts or performance by women's age. Generally, for all women, screening metrics improved when DBT was used as an adjunct to FFDM. Results relating to breast density are complicated by the limitations of the breast density classification systems and the comparators used in the existing research. Age-related comparisons are limited by different age groupings reported in the primary studies.

Question 4: What are the incremental costs associated with implementing DBT as a screening tool (including alone, with s2DM, or as an adjunct to FFDM) for the early detection of breast cancer compared to FFDM alone?

There is limited information about the actual costs associated with either the implementation or use of DBT in a screening setting. Some available articles describe the kinds of costs that may be incurred with the use of DBT (such as capital upgrade expenditure, people resourcing costs, etc.) but do not provide financial estimates. Nor do these articles cover the range of possible ways that DBT could be integrated (either as an adjunct screen, with s2DM or alone). Other papers provide modelled analysis estimations for implementation, but these are based on the US (which operates a different health funding model compared to the Australian system and is

therefore of limited usefulness). These models indicated some evidence that DBT in combination with FFDM results in overall economic favourability, when balanced against the potential improved clinical outcomes. More information about the cost of implementation in the Australian context is needed.

Question 5: Do asymptomatic women screened for breast cancer experience more anxiety, discomfort or inconvenience if the screening strategy is DBT alone, DBT + s2DM or FFDM + DBT compared to FFDM alone?

There is insufficient evidence describing women or practitioners' acceptability of DBT as a screening method. Data from practical evaluations, inferences from participant selection of screening test, and intuition suggest that women appreciate the lower compression that may accompany DBT (depending on screening strategy used) and the reduced anxiety associated with reduced recall and false positive rates. More research that tests and validates women's experiences is needed.

Assessment of evidence table summary

Table 1A: Assessment of evidence for FFDM + DBT from all studies

Outcomes	Participants Studies Follow up	Quality of evidence	Overall results
Reduction in mortality	0	Nil	No studies reported on a reduction in mortality.
Cancer detection rate	1,009,427 participants 20 studies	⊕⊕⊕⊕ Strong	Pooled analysis of data from two prospective, fully paired studies embedded in population-based screening programs: FFDM + DBT increases CDR by 2.43 cancers per 1000 screening examinations. A third prospective, fully paired study also reported increased CDR of 2.2 cancers per 1000 screening examinations. Data from 15 other retrospective studies of different design and variable quality report increased incremental CDR (although few studies reported statistically significant results).
Invasive cancer detection rate	881,525 participants 10 studies	⊕⊕⊕⊕ Strong	Pooled analysis of data from two prospective, fully paired studies embedded in population-based screening programs: FFDM + DBT increases invasive CDR by 2.33 cancers per 1000 screening examinations. Data from 7 other retrospective studies of different design and variable quality report increased incremental CDR.
Interval cancer detection	55,457 participants 3 studies	⊕ Very low	Few studies report interval cancer rates. No systematic review or pooled analysis was available.
PPV_{1-3}	1,347,022 participants 13 studies	⊕⊕ Low	No systematic review or pooled analysis was available. Data from one fully paired trial reported a small non-statistically significant increase in PPV. 12 studies reported increased PPV favouring FFDM + DBT compared to FFDM alone.
Recall rate	572,555 participants 15 studies	⊕⊕ Low	Pooled analysis and data from prospective fully paired trials and retrospective analysis shows inconsistent results.
False positive rate	217,565 participants 6 studies	⊕⊕ Low	Pooled analysis and data from prospective fully paired trials and retrospective analysis shows inconsistent results.
Radiation dose	18,926 participants 4 studies	⊕⊕ Low	No systematic review or pooled analysis was available.
Interpretation time	20,178 participants 5 studies	⊕ Very low	No systematic review or pooled analysis was available.

Table 2B: Assessment of evidence for FFDM + DBT from one RCT and three fully-paired trials

Outcomes	Participants Studies Follow up	Quality of evidence	Overall results
Reduction in mortality	0	Nil	No studies reported on a reduction in mortality.
Cancer detection rate	30,822 participants 4 studies	⊕⊕⊕⊕ Strong	Pooled analysis of data from two prospective, fully paired studies embedded in population-based screening programs: FFDM + DBT increases CDR by 2.43 cancers per 1000 screening examinations. A third prospective, fully paired study also reported increased CDR of 2.2 cancers per 1000 screening examinations.
Invasive cancer detection rate	19,923 participants 2 studies	⊕⊕⊕⊕ Strong	Pooled analysis of data from two prospective, fully paired studies embedded in population-based screening programs: FFDM + DBT increases invasive CDR by 2.33 cancers per 1000 screening examinations.
Interval cancer detection	10,889 participants 2 studies	⊕ Very low	A total of eight interval cancers were detected by these two studies but there is insufficient data to calculate to complete further analysis.
PPV_{1-3}	12,631 participants 1 study	⊕ Very low	A non-significant result was reported with FFDM + DBT having a slightly higher PPV than FFDM alone.
Recall rate	30,822 participants 4 studies	⊕⊕ Low	Pooled analysis and data from prospective fully paired trials shows inconsistent results.
False positive rate	30,822 participants 4 studies	⊕⊕ Low	Data from prospective fully paired trials shows inconsistent results.
Radiation dose	12,631 participants 1 study	⊕ Very low	MGD is higher with FFDM + DBT compared to FFDM alone.
Interpretation time	19,923 participants 2 studies	⊕ Very low	Reading time doubled.

Table 3A: Assessment of evidence for DBT + s2DM from all studies

Outcomes	Participants Studies Follow up	Quality of the evidence	Overall results
Reduction in mortality	0	Nil	No studies reported on a reduction in mortality.
Cancer detection rate	153,668 participants 5 studies	⊕⊕ Low	No systematic review or pooled analysis was available.
Invasive cancer detection rate	131,726 participants 3 studies	⊕ Very low	No systematic review or pooled analysis was available. Results from individual studies reported that CDR increased for DBT + s2DM compared to FFDM in the results from two prospective, fully paired studies. Results varied for DBT + s2DM compared to FFDM + DBT.
Interval cancer detection	NA	Not reported	No data reported.
PPV_{1-3}	162,718 participants 4 studies	⊕⊕ Low	No systematic review or pooled analysis was available. Data from one fully paired trial reported a small non-statistically significant increase in PPV. 12 studies reported increased PPV favouring FFDM + DBT compared to FFDM alone.
Recall rate	148,814 participants 4 studies	⊕⊕ Low	Pooled analysis and data from prospective fully paired trials and retrospective analysis shows inconsistent results.
False positive rate	100,752 participants 3 studies	⊕⊕ Low	Pooled analysis and data from prospective fully paired trials and retrospective analysis shows inconsistent results.
Radiation dose	53,198 participants 5 studies	⊕⊕⊕ Moderate	No systematic review or pooled analysis was available but consistent findings across a range of study types (including manufacturer data) indicates DBT + s2DM radiation dose is considerably less than that used to acquire images with FFDM.
Interpretation time	31,254 participants 7 studies	⊕ Very low	No systematic review or pooled analysis was available.

Table 4B: Assessment of evidence for DBT + s2DM from two fully paired trials

Outcomes	Participants Studies Follow up	Quality of the evidence	Overall results
Reduction in mortality	0	Nil	No studies reported on a reduction in mortality.
Cancer detection rate	21,942 participants 2 studies	⊕ Very low	Results from the two studies reported mixed results: one found higher cancer detection rates with DBT + s2DM (p<.0001); the other found higher rates with FFDM + DBT compared to DBT + s2DM (no significance testing or 95%CI provided).
Invasive cancer detection rate	0	Not reported	No studies reported on invasive cancer detection rate.
Interval cancer detection	0	Not reported	No studies reported on interval cancer detection.
$PPV_{1\text{-}3}$	0	Not reported	No studies reported on $PPV_{1\text{-}3}$.
Recall rate	0	Not reported	No studies reported on overall recall rate.
False positive rate	21,942 participants 2 studies	⊕⊕ Low	Data from prospective fully paired trials shows inconsistent results.
Radiation dose	21,942 participants 2 studies	⊕⊕ Low	MGD is higher with FFDM + DBT compared to DBT alone or FFDM alone.
Interpretation time	0	Not reported	No studies reported on reading time for this imaging combination.

Table 5A: Assessment of evidence for DBT_{MLO} compared to other imaging combinations from all studies

Outcomes	Participants Studies Follow up	Quality of the evidence	Overall results
Reduction in mortality	0	Nil	No studies reported on a reduction in mortality.
Cancer detection rate	7,681 2	⊕ Very low	No systematic review or pooled analysis was available.
Invasive cancer detection rate	7,681 2	⊕ Very low	No systematic review or pooled analysis was available.
Interval cancer detection	0	Nil	No data reported.
PPV_{1-3}	0	Nil	Not reported
Recall rate	7,681 2	⊕ Very low	No systematic review or pooled analysis was available.
False positive rate	7,681 2	⊕ Very low	No systematic review or pooled analysis was available.
Radiation dose	0	Not reported	No systematic review or pooled analysis was available.
Interpretation time	7,681 2	⊕ Very low	No systematic review or pooled analysis was available.

Table 6B: Assessment of evidence for DBT_{MLO} compared to other imaging combinations from one fully-paired trial

Outcomes	Participants Studies Follow up	Quality of the evidence	Overall results
Reduction in mortality	0	Nil	No studies reported on a reduction in mortality.
Cancer detection rate	7500 participants 1 study	⊕ Very low	DBT_{MLO} + DM_{CC}: 8.9 cancers detected per 1000 screening examinations FFDM: 6.3 cancers detected per 1000 screening examinations
Invasive cancer detection	0	Nil	No studies reported on invasive cancer detection.
Interval cancer detection	0	Nil	No studies reported on interval cancer detection.
PPV_{1-3}	0	Nil	No studies reported on PPV_{1-3}.
Recall rate	7500 participants 1 study	⊕ Very low	DBT_{MLO} + DM_{CC}: 3.8% FFDM: 2.6% (but note very low overall recall rate in this screening program)
False positive rate	7500 participants 1 study	⊕ Very low	DBT_{MLO}: 1.7% DM_{CC}: 0.9% FFDM: 1.1% (but note very low overall recall rate in this screening program)
Radiation dose	0	Not reported	No data was available.
Interpretation time	7500 participants 1 study	⊕ Very low	Reading time doubled.

DASHBOARD

Table 7: Current evidence base

The current evidence base comparing tomosynthesis to digital mammography reports that …			
		FFDM + DBT	**DBT + s2DM**
Fast	Image acquisition time	Between 3 and 25 seconds to capture DBT image, which is time on top of FFDM image acquisition	Between 3 and 25 seconds to capture both DBT and 2D images
	Compression time	Up to one minute longer than FFDM alone if using Hologic's Dimensions system	Does not use FFDM so faster than dual acquisition and may be able to acquire acceptable images with lower compression
	Interpretation time	Depends on radiologist experience but is generally more than 25% longer than that required to review FFDM images alone	Depends on radiologist experience but is generally more than 25% longer than that required to review FFDM images alone
Sensitive	Reduction in mortality	No evidence available	No evidence available
	Cancer detection rate	Strong evidence that DBT as an adjunct screen to FFDM increases in CDR More evidence on the association between CDR across different screening strategies is needed	Low (but emerging) evidence of an increase in CDR screening examinations compared to FFDM alone Emerging evidence of comparable CDR compared to FFDM + DBT
	PPV	Low evidence of an increase in PPV_{1-3} with DBT compared to FFDM alone	Low evidence of an increase in PPV_1 compared to both FFDM + DBT and FFDM alone. Limited evidence of an increase in PPV_{2+3}
	Sensitivity	More studies with longer-term follow-up are needed to determine this	More studies with longer-term follow-up are needed to determine this
	Interval cancer	More studies with longer-term follow-up are needed to determine this	No evidence available
Specific	Recall rate	Some evidence that DBT as an adjunct to FFDM decreases recall rates. More evidence to confirm this finding is needed.	Some evidence that DBT + s2DM decreases recall rates. More evidence to confirm this finding is needed.
	False positive	Some evidence that DBT as an adjunct to FFDM reduces false positive recall rates. More evidence to confirm this finding is needed.	Some evidence that DBT + s2DM decreases recall rates. More evidence to confirm this finding is needed.
Radiation dose	Within BSA accreditation standards	MGD is almost double compared to FFDM but is still within acceptable limits	MGD is 20% lower than FFDM alone and 50% lower than FFDM + DBT but further information on image quality is needed

The current evidence base comparing tomosynthesis to digital mammography reports that …			
		FFDM + DBT	**DBT + s2DM**
Acceptability	To women	More evidence validating women's experiences is needed	More evidence validating women's experiences is needed
	To clinicians and health practitioners	More evidence validating practitioners' experience is needed	More evidence validating practitioners' experiences is needed
Financial implications	To women and health systems	Some evidence of cost effectiveness but more cost analysis is needed.	More cost analysis is needed.
Cost-effectiveness		Limited available evidence	Limited available evidence

1. INTRODUCTION

1.1. About digital breast tomosynthesis

Digital breast tomosynthesis (DBT) (also known as breast tomosynthesis, mammographic tomosynthesis or three dimensional/3D mammography) is an imaging technology that can be used to detect, assess and diagnose breast cancer. DBT records between 11 and 25 low-dose images of a compressed breast depending on the imaging system used[3]. These images are reconstructed in 1mm parallel slices to form a three-dimensional image of the breast. Radiologists (or other readers) then analyse these images to determine the presence of suspected abnormalities or to further investigate an area identified as suspicious on a digital mammogram. The thin cross-sectional images created by DBT minimise the masking effects of breast tissue overlap, which can improve margin visibility for soft tissue tumours and increase lesion conspicuity. This potentially increases screening sensitivity and specificity (especially for women with dense/non-fatty breasts) as abnormalities are easier to see.

Radiation dose varies depending on whether DBT is used alone, with integrated s2DM image acquisition (s2DM[4]) or is used as an adjunct to full field digital mammography (FFDM[5]). We know that FFDM + DBT requires a higher radiation dose than FFDM alone to acquire images during a screening examination. Concerns about radiation dose plus the longer image acquisition and interpretation time required with FFDM + DBT means that this screening strategy could be potentially unacceptable to women and practitioners. DBT + s2DM developed in response to these concerns. As a result, DBT's use (both in clinical and research settings) is evolving as is the evidence base underpinning the use of DBT + s2DM within a screening program grows.

DBT (using Hologic's Dimensions system) was first approved for use in breast cancer screening by the FDA in 2011[6]. Since then, other systems capable of performing DBT have also been approved. DBT is not widely used as a primary screening tool within any national breast-screening program although some European jurisdictions including France and Monaco include DBT as a screening option (Liberatore et al., 2017). It is not used as a primary test for average-risk women in the BreastScreen Australia (BSA) program; however, access to DBT imaging for screening purposes is offered through some private radiology clinic settings. Outside the BSA program, DBT is increasingly used for the assessment of both screen-detected abnormalities and symptomatic breast cancers.

1.2. BreastScreen Australia's position statement on tomosynthesis

In 2014, the Community Care and Population Health Principal Committee of the Australian Health Ministers' Advisory Council endorsed BSA's position statement on DBT. This position statement was based on a literature review completed in 2009 (Department of Health and Ageing, 2009) and other papers published between 2009-13.

---

[3] Hologic's Dimensions system takes 15 projections taken over approximately 4 seconds. Other CE mark or FDA-approved systems use 9 or 25 projections taken over 3 to 25 seconds (Sechopoulos, 2013).
[4] s2DM is a two-dimensional mammogram that is generated from a DBT source data. These reconstructed images are like those captured in the mediolateral (MLO) and craniocaudal (CC) views used in a standard FFDM screening examination (Freer et al., 2017).
[5] FFDM is also known as two-view digital mammography.
[6] https://www.accessdata.fda.gov/cdrh_docs/pdf8/P080003b.pdf. Accessed 10 February 2018.

The BSA position statement on DBT says that it:

> *"has the potential to decrease the number of women who are recalled for further tests (reduce recall rates) and possibly increase the detection of breast cancer (improve sensitivity).";* however, the balance between *"relative harms and benefits to well women of radiation dose, and the cost, efficiency and effectiveness of using this technology are as yet unclear".*

The Standing Committee on Screening concluded that FFDM remained the most effective population screening technology for breast cancer.

Since publication of BSA's position statement, the evidence base for DBT as a promising effective population screening tool for the early detection of breast cancer in asymptomatic women has continued to develop. At the same time, evolution in the way that DBT is used continues (including new clinical and research protocols involving DBT as a stand-alone screening strategy or DBT with s2DM). Emerging evidence includes further peer-reviewed research from large prospective trials investigating DBT's effectiveness as a primary screening tool and its safety, changes to screening strategy using DBT which affect the radiation dose, cost-effectiveness and financial considerations. More information is also becoming available on the practical considerations that require thought if DBT is to be introduced as a screening technology. More population-based trials are underway, one of which reported interim findings with further findings expected in 2018 (Lång et al., 2016A).

1.3. Purpose and scope of this literature review

The Department of Health engaged Allen and Clarke Policy and Regulatory Specialists Limited (*Allen + Clarke*) to:

- complete a literature review of the evidence base informing the BSA position statement on DBT as a screening strategy for the early detection of breast cancer in asymptomatic women aged over 40 years, and
- prepare any updates to the BSA's position statement on DBT if considered necessary by the Breast Screening Technical Reference Group.

We wanted to know if DBT (implemented either alone, with s2DM or as an adjunct to FFDM) is a more sensitive, specific and safer test for the early detection of breast cancer in asymptomatic women compared to bilateral FFDM alone (the current 'gold standard'). The review of evidence included considering if available published evidence on efficacy, effectiveness and safety indicates that DBT should be the preferred method for screening asymptomatic women for breast cancer in Australia (i.e., that it replaces or is used as an adjunct to FFDM for all or some women). We were also interested in whether DBT should be the preferred method for average risk women and/or population groups with a higher than average lifetime risk of breast cancer. We also explored the incremental costs associated with implementing DBT as a screening tool and women's experience with this imaging technology when used for screening purposes. Initially, the answers to these questions will support the Breast Screening Technical Reference Group's consideration of what updates (if any) are needed to BSA's position statement on DBT.

The literature review is not a systematic review. It also did not explore the role of DBT in the assessment of suspected or symptomatic breast cancer or diagnosis of breast cancer beyond the screening pathway: it focuses on DBT's role in screening asymptomatic women to the point of confirmation of the presence or absence of an abnormality. The Department has commissioned

a separate literature review on the role of DBT in the assessment of screen-detected or symptomatic breast abnormalities. It also did not consider other technologies that may have a population screening application for the early detection of breast cancer in asymptomatic women who have an average lifetime risk of breast cancer or investigate the relationship between DBT and screening technologies other than modern FFDM (as this is accepted as the most effective population screening technology currently used in Australia). Finally, no original meta-analysis or other pooled analysis was completed.

1.4. Ongoing research

Our review of www.clinicaltrials.gov (completed on 8 January 2018) identified three large active studies investigating the role of DBT in population-based screening for asymptomatic women (either alone or in combination with s2DM). Further studies are recruiting study participants. Further updates to the BSA position statement may be required as these studies report interim or final findings. Key ongoing or upcoming trials include the:

- Malmö Breast Tomosynthesis Screening Trial (a single site study in Sweden)[7]
- Tomosynthesis Trial in Bergen (a single site study in Norway)[8], and
- Tomosynthesis Mammographic Imaging Screening Trial (a multi-armed, multicentre study in Canada)[9].

Other large studies (either planned or in recruitment) include the PROSPECTS trial in the United Kingdom[10], the TOSYMA study in Germany[11] and two large Italian studies[12]. The Maroondah trial (Australia) will provide relevant information about DBT's implementation in a screening setting.

---

[7] This clinical trial of 15,000 participants investigates if more breast cancers can be detected with DBT_{MLO} compared to DM_{CC} or FFDM in a population invited to screening. Interim results have been presented and are discussed in this report (Lång et al., 2016A, 2016B). The primary completion date is 31 December 2017. Further data will be made available when final study data is published.

[8] This prospective cohort study of 29,453 women will compare DBT + s2DM to digital mammography as a screening tool for women aged 50-69 years participating in a population-based screening program. Study outcomes focus on cancer detection, interval cancer rates, PPV, recall rates, prognostic and predictive tumour characteristics, radiation dose, interpretation time, and cost-effectiveness. This study began in January 2016 and primary study completion is set for January 2020.

[9] This is a randomised trial of up to 164,946 women to compare diagnostic accuracy of screening for breast cancer with FFDM + DBT compared to FFDM alone. Key outcomes include cancer detection, recall rates, interval cancers, prevalence of breast cancer subtypes, clinical characteristics of cancers, radiation dose, observer performance studies, BIRADS imaging features, breast cancer mortality, quality monitoring outcomes, PPV, health care utilisation, false positives/true negatives, biopsy rates and biomarker correlation. The primary lead-in study completion date is November 2018. The final study completion date is 2030.

[10] The PROSPECTS RCT has a proposed sample of 100,000 women to investigate the cost-effectiveness of breast cancer screening using FFDM + DBT compared to DBT + s2DM. It aims to demonstrate that the DBT is not inferior to FFDM + DBT. It will report on recall rates, benign biopsy rates at diagnostic assessment and surgery; measure the effect of FFDM + DBT compared to FFDM for groups by age, breast density and prevalent/incident screen; develop methods to measure reader performance and complete reader studies; compare women's preferences for screening strategy. The RCT will happen over seven years from 2018 with initial results to be presented within 18-24 months.

[11] This prospective, randomised trial of 80,000 asymptomatic women is due to begin recruiting in March 2018. It will investigate DBT + s2DM compared to FFDM. Study completion is proposed for March 2023. Key outcome measures focus on cancer detection rates including by cancer type and category, interval cancer rate, recall rate and PPV.

[12] One RCT study of 40,000 women will compare DBT + s2DM to FFDM. It will investigate interval cancer, recall, PPV, biopsy rates, cancer detection and self-reported pain/discomfort during mammography. The other RCT of 92,000 asymptomatic women will investigate FFDM + DBT compared to FFDM in a multicentre, population screening program and will investigate cancer detection, interval cancer, recall rate, PPV, and false positives. The first study will be completed in December 2018. The second study will be completed in December 2019.

1.5. Imaging systems used in studies reported in this literature review

We assessed the imaging systems and screening strategies used in all studies included in our review of research on screening clinical outcomes and performance metrics like cancer detection rate, sensitivity, PPV, specificity and radiation dose.

1.5.1. Imaging systems used in the literature

There is a very high degree of homogeneity between the DBT-capable imaging systems used in the studies informing this review. Almost all studies used Hologic's Selenia Dimensions for all DBT imaging. Rodriguez-Ruiz et al. (2017) and Lång et al. (2016A) used Siemens Mammomat Inspirations for DBT. Starikov et al. (2016) did not describe the imaging system used. Most studies also used Hologic systems for FFDM imaging. The remaining studies used other systems to generate digital mammography images including GE's SenoGraphe Essential, Senographe 2000D, Senographe DS, or Fuji CRM.

This literature review includes five studies reporting on DBT + s2DM in breast cancer screening. Four studies used Hologic's Selenia Dimensions system and C-view 2D software for digital mammography image acquisition (which became available following FDA-approval in 2013). The other study also used Hologic's Selenia Dimensions but did not report on which software was used for the digital mammography image reconstruction.

There is limited evidence of the reproducibility of the results presented if other systems are used (although other DBT-capable systems have been developed including GE SenoClaire, FujiFilm ASPIRE Cristalle, and Siemens Mammomat Inspirations). These systems may differ to the Hologic system by imaging geometry, angular range, number of projections, scan duration, acquisition method, detector technology, and reconstruction algorithms. Vedantham et al. (2015) provide more detailed discussion of the differences and similarities.

1.5.2. Screening strategies used in the literature

DBT can be used alone or as an adjunct to FFDM to create a sensitive screening tool. As noted by Coop et al. (2016), the research base includes a range of comparisons of different screening strategies which have been used on different groups of women (including both those of average-lifetime risk of breast cancer and those with higher risk, younger women versus older women). Different implementation strategies including reading strategies have also been assessed and compared. The different combinations of screening strategy reported in the literature are as follows:

- Most research focused on the use of FFDM + DBT compared to FFDM alone: 43 articles generated from 35 studies reported on screening metrics based on comparison between FFDM + DBT and FFDM alone.
- The next most common pairing under research investigation is DBT + s2DM compared to FFDM + DBT and/or FFDM alone (*n*=5).
- A small number of studies (*n*=2) which explore CDR for single-view screening strategies including DBT_{MLO} or DM_{CC}.

2. METHODOLOGY

Summary

- This literature review provides an overview of research about the effectiveness and safety of DBT as a population screening tool for the early detection of breast cancer in asymptomatic women. It is not a systematic review. We have provided statements about the quality of the evidence included in this review. No primary research or pooled analysis was undertaken.
- The following databases were searched on 15 and 19 December 2017: EMBASE, Ovid Medline, CINAHL, ProQuest and Scopus. The following websites were reviewed: clinicaltrials.gov, the Cochrane database, NICE, INAHTA, and the UK NHSBPS.
- All returned citations and abstracts were assessed for relevance to the research questions and inclusion criteria. The same criteria were used to review the full-text and bibliographies of all articles proposed for inclusion. The methodologies of all included studies were critically appraised using the AMSTAR 2 tool or SIGN criteria.
- A total of 85 articles met the inclusion criteria.

2.1. Objectives

This literature review explores if FFDM remains the best test for the early detection of breast cancer in asymptomatic women or if DBT (either alone, with s2DM or as an adjunct to FFDM) is more sensitive, specific and safer and therefore better at detecting breast cancer early. The effectiveness of DBT as a screening tool for specific population groups (including women with dense breast and younger women) and implementation considerations such as financial costs and workflow are explored. A systematic review with pooled analysis was not performed.

2.2. Research questions

2.2.1. Questions about effectiveness, efficaciousness and safety

The three questions about effectiveness, efficaciousness and safety were:

1. Should DBT (with or without s2DM) be used as the primary screening tool for the early detection of breast cancer in asymptomatic women aged over 40 years?
2. For the early detection of breast cancer in asymptomatic women aged over 40 years, is DBT (including s2DM):
   - a more efficacious and safer screening modality than FFDM alone?
   - in addition to FFDM, a more efficacious and safer screening test than FFDM alone?
3. For the early detection of breast cancer, are there population groups for which DBT (including s2DM):
   - is a more efficacious, safer screening modality than FFDM alone?
   - in addition to FFDM is a more efficacious and safer screening modality than FFDM alone?

The PICO(S) criteria underpinning these research questions are described in Table 5 (below).

Table 8: PICO(S) criteria for questions relating to effectiveness and safety

Criterion	Description
Populations	Women aged over 40 years with no symptoms of breast cancer Women living in rural or remote communities Women with dense/non-fatty breasts Ethnic groups including Aboriginal and Torres Straits Islanders Women with other risk factors for breast cancer including familial history or previous history of breast cancer
Intervention	DBT (either alone or when combined with s2DM)
Comparators	FFDM alone FFDM when used in combination with DBT
Outcomes	Radiation dose by combination of screening modality Screen detected cancer rates Sensitivity in detecting cancers present (detection rate for types/sub-types of breast lesions) Specificity (recall rates, false-positive recall rates and over-diagnosis for specific types/sub-types of breast lesions) Interval cancer rates Surrogate mortality indicators (tumour size at detection, lymph node negativity, grade)
Study types	Systematic reviews, randomised controlled trials (RCT)

2.2.2. Question about implementation

The question on implementation was:

What are the incremental costs associated with implementing:

- DBT (including s2DM) as a screening tool for the early detection of breast cancer compared to FFDM alone?
- DBT (including s2DM) plus FFDM as a screening tool for the early detection of breast cancer compared to FFDM alone?

The PICO(S) criteria underpinning these research questions are described in Table 6 (overleaf).

Table 9: PICO(S) criteria for questions relating to the implementation of DBT

Criterion	Description
Population	Women aged over 40 years (inclusive) with no symptoms of breast cancer Women aged over 40 years (inclusive) with no symptoms of breast cancer with dense breasts
Intervention	DBT (combined with s2DM or used alone)
Comparators	FFDM alone FFDM when used in combination with DBT
Outcomes	Work flow benefits including imaging acquisition time and interpretation/reading time Technologist and radiologist training IT changes (including software/hardware upgrades and data storage)
Study types	Systematic reviews, RCT, observational studies, health technology assessments, grey literature

2.2.3. Question about acceptability to women

The question on acceptability was:

> Do asymptomatic women screened for breast cancer experience more anxiety, discomfort or inconvenience if the screening modality is:
>
> - DBT (including s2DM) compared with FFDM alone?
> - DBT (including s2DM) plus FFDM compared with FFDM alone?

The PICO(S) criteria underpinning these research questions are described in Table 7 (below).

Table 10: PICO(S) criteria for questions relating to the acceptability of DBT

Criterion	Description
Population	Women aged over 40 years (inclusive) with no symptoms of breast cancer Women living in rural or remote communities
Intervention	DBT (combined with s2DM or used alone)
Comparators	FFDM alone FFDM when used in combination with DBT
Outcomes	Discomfort/pain Anxiety/distress Time spent having a mammogram/convenience Women's confidence in screening modality
Study types	Systematic reviews, RCT, observational studies, HTA, grey literature

2.3. Literature search

The following databases were searched on 15 and 19 December 2017:

- CINAHL
- Clinicaltrials.gov
- Cochrane Library database
- Embase
- National Institute for Health and Clinical Excellence
- OVID Medline
- ProQuest
- Scopus
- UK National Institute for Health Research HTA database, and
- UK NHSBPS.

To complete a systematic search, we used combinations of subject/index terms where appropriate (eg, exploded term 'mammography' or exploded 'breast neoplasm') in combination with key words, or key words alone depending on the search functionality of each database or website (eg, main searches included 'tomosynthesis' PLUS 'breast cancer' PLUS 'screen*' in the title or abstract).

The following limits were applied on all searches:

- a date criterion (1 January 2010 – 31 December 2017 or 2010 onwards)
- English language, and
- study type restrictions (where available and appropriate, we restricted returns from research databases to peer-reviewed systematic reviews, literature reviews, RCT, observational studies and clinical trials).

'Human' was not used as a limiter as no animal studies were returned for the search terms. Duplicate citations and a small number of false hits/inaccurate returns were removed before all initial returned citations and abstracts were reviewed for relevance to the main research questions. Material was excluded if it:

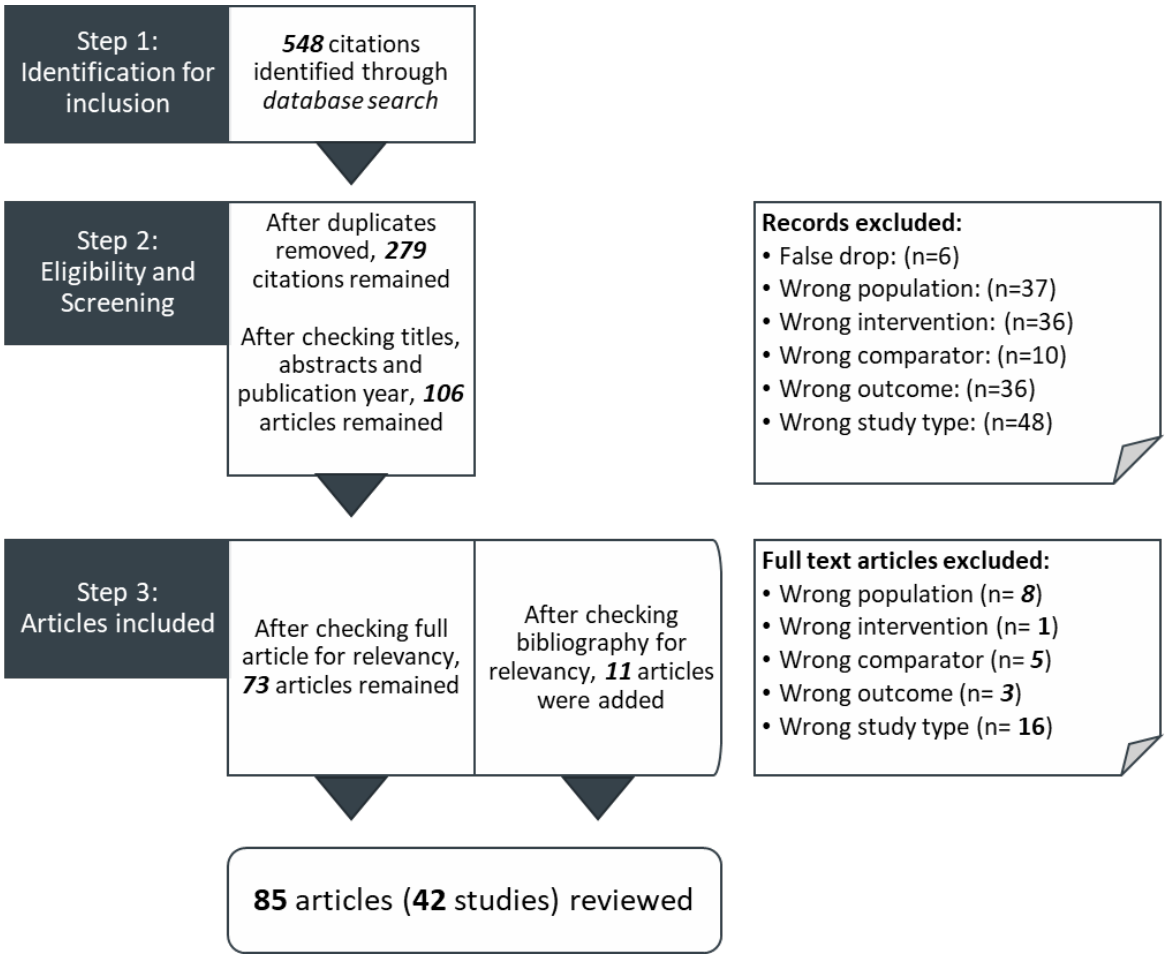
- did not relate to DBT as a population screening tool for breast cancer (i.e., if it related to the role of DBT in the diagnosis or assessment of breast cancer)
- compared DBT to screening strategies other than FFDM
- focused on a study population other than asymptomatic women
- related to surveillance.

To determine if this first search retrieved the correct range of available research, a validation process was completed using five recent systematic or literature reviews relevant to the primary research questions (Houssami, in press; Skaane, 2017; Coop et al., 2016; Hodgson et al., 2016; Houssami & Turner., 2016). There was a high degree of consistency between in the studies returned using our strategies and those included in the three reviews.

From this first sweep, full texts for all proposed inclusions were retrieved and reviewed for relevance to the research questions, inclusion criteria and documented PICOT criteria. A critical appraisal of study design (to determine overall quality) was completed and the bibliography of each included article was reviewed to identify other relevant research that may be of interest.

The citation review process for academic articles relating to the research questions is described in Figure 1 (below).

Figure 1: Citations review process



Study types were:

- Two systematic reviews
- 12 narrative literature reviews
- One randomised controlled trial
- 56 articles
- Nine position statements
- Four technical or practical evaluations of DBT systems, and
- One commentary.

2.4. Limitations and interpretation

The evidence base underpinning DBT as a screening strategy and evidence on the most effective and safest screening strategy continues to develop rapidly. Further large population-based trials are underway (including those due to report final findings in 2018). In addition, a systematic review and meta-analysis is underway exploring whether imaging with DBT improves cancer detection, recall rates and incremental cancer detection rates compared to digital mammography. As such, the evidence for DBT as a primary screening tool and advice on preferred screening strategy may strengthen over the next 12 months. Updating this literature review in due course will be necessary.

Much of the evidence about the detection, sensitivity and specificity of DBT compared to FFDM published is based on screening strategies that compare FFDM + DBT to FFDM alone. More recent research (including some large, active population-based trials like the Bergen, PROSPECTS, TOSYMA and two Italian trials) will focus on a screening strategy using DBT alone or with s2DM images compared to FFDM + DBT or FFDM alone. As such, it is likely that the currently available evidence is not keeping pace with the way that DBT is being used in clinical or research settings. Given that the DBT + s2DM has a lower radiation dose compared to an integrated FFDM + DBT screening strategy, it is likely that the evidence may underestimate the safety of DBT as a screening tool. Further evidence will clarify this issue.

Inclusion criteria for Research Questions 1 – 3 specified that only systematic reviews and RCTs would be included. Much of the available evidence published since 2010 is based on evidence from three large prospective trials or retrospective observational research design (either from observer performance studies or single or multi-site-analysis). The published literature includes literature summaries (including Houssami in press; Poplack, 2017; Skaane, 2017; Vedantham et al., 2015); however, our literature review identified only one RCT and two systematic reviews (one of which was of lower quality). Only one of the systematic reviews (Hodgson et al.) included pooled analysis (and they only included data from two prospective trials in this analysis due to the heterogeneity in study design of the other studies investigated). To ensure that we considered an adequate range of published literature, we therefore expanded the study type criteria to include primary data from observational studies. Primary studies already incorporated into systematic or literature reviews (Houssami in press; Coop et al., 2016 Hodgson et al., 2016; Houssami & Turner, 2016; Vedantham et al., 2015) were reviewed but not separately assessed unless additional material not described in the systematic or literature review was included. Relevant data from all primary studies is included in evidence tables. While narrative and non-systematic reviews may lack some essential components (such as: clear eligibility criteria, search strategies, study selection processes, outcomes, and assessment of bias in individual studies), and that inclusion of narrative literature reviews may introduce reporting bias, inclusion is warranted as this provides a bigger picture view and a wider look at the current evidence available related to DBT.

GRADE assessment was undertaken as a measure of the strength of evidence; however, this literature review found only one small RCT with a focus on one sub-population of interest to our research questions (women aged 40 to 49 years). This impacts on the application of the GRADE methodology. Our literature does report on three large prospective trials embedded in a population-based breast cancer screening program, each of which used a fully paired study design. Each woman participating in the trial was imaged with both FFDM and the DBT screening strategy under investigation (eg, FFDM + DBT, DBT + s2DM, or DBT_{MLO}). This means that each woman was effectively her our control (FFDM imaging) and case (DBT imaging). We

have determined that data from the STORM, OTS and Malmö trials can be used in GRADE in a similar fashion to that expected of an RCT.

Evidence from longer-term prospective trials involving sufficient numbers of asymptomatic women to detect small changes in cancer detection rates (CDR) is being published. Because long-term, adequately powered studies are limited, currently available data about DBT's impact on interval cancer rate, mortality benefits and improvement in treatment morbidity compared to FFDM is limited. Proxy measures (such as tumour size at detection) are available but these do not provide a sense of the long-term mortality reduction benefits which DBT may offer compared to FFDM alone. Again, future research will provide more information that can be used to assess this.

The Department of Health has commissioned another literature review to investigate DBT's role in assessment and diagnosis. Studies focused on diagnostic populations have therefore been excluded from this review but will be explored in the next. Also, information about breast density and the role of supplemental screening is further explored in *Allen + Clarke*'s literature review on breast density.

3. EFFECTIVENESS AND SAFETY OF DBT AS A SCREENING TOOL

We want to know, based on current evidence, what role DBT could play in a modern breast cancer screening program. We want to know, based on current evidence, if DBT (either alone or integrated with other mammography screening strategies) can safely detect breast cancers that are present (even if small or asymptomatic). Chapter 3 investigates the effectiveness and safety of DBT in a screening environment. Three main screening strategies, which reflect research focus, are assessed:

1. DBT as an adjunct screening tool (i.e., FFDM + DBT) compared to FFDM alone
2. DBT + s2DM compared to FFDM + DBT or FFDM alone, and
3. One-view DBT (DBT_{MLO}) plus or compared to one-view digital mammography (DM_{CC}) or FFDM.

The results are presented by the following clinical outcomes and performance metrics:

- sensitivity (i.e., CDR, invasive CDR and tumour characteristics, interval cancer rates)
- positive predictive value (PPV)
- specificity (i.e., overall recall rate and false positive rate)
- safety (i.e., radiation dose), and
- the impact DBT may have on women with heterogeneously dense or extremely dense breasts or by age.

Chapter 4 discusses implementation including the impact of DBT on reader performance, which is affected by both a program's reading strategy (single or double) and individual reader experience.

The discussion for each outcome includes a description of the number of studies identified, the overall quality of the studies, and a summary of the results from all studies. Detailed study tables provide additional material about study population, methodology, intervention, comparator and key results. This information is used to answer the three questions about effectiveness, efficaciousness and safety and provide a statement of the quality of the evidence underpinning the answers to these questions.

3.1. Sensitivity

Sensitivity (the proportion of asymptomatic breast cancers correctly identified by a screening test, or the true positive/negative rate) is an important dimension of an effective population-based breast screening program. Clinical outcomes that can impact on sensitivity are CDR, interval cancer rate and tumour characteristics at detection.

Key findings

There is strong evidence that DBT increases all cancer detection rates (CDR) and invasive CDR compared to FFDM alone. All studies with adequate statistical power to detect small changes to CDR reported increased cancer detection. These increases were seen in studies with a range of designs and for different combinations of DBT and digital mammography. Increases were seen in all DBT screening strategies used: FFDM + DBT compared to FFDM alone, DBT + s2DM compared to FFDM alone, and DBTMLO compared to either DMCC or FFDM alone. Results from large, fully paired prospective trials embedded in population screening programs consistently reported statistically significant incremental cancer detection rates of more than 2 per 1000 screening examinations. That is, the use of DBT resulted in cancer detection in at least two more women compared to FFDM alone. Smaller (possibly underpowered) studies also reported the same increase in CDR but the incremental increases reported were smaller. Smaller increases may also be due to annual rather than biennial screening or single compared to double reading strategies. Further, DBT + s2DM has cancer detection rates that are at least as good as FFDM + DBT but this approach delivers a reduced radiation dose. DBT + s2DM is a very promising approach.

Most of the studies investigating cancer detection rate (CDR) have short time frames (<24 months). Further research is needed to determine the overall mortality and treatment morbidity benefit conferred by DBT as a screening strategy in a population-based screening environment (including further evidence about interval cancer rate) and if the improvement in cancer detection rates provided by DBT (either with integrated s2DM or with FFDM) is sustained between first and subsequent (i.e., prevalent and incident) screening examinations.

This literature review describes CDR findings including pooled analysis from one systematic review and one literature review. Primary studies already incorporated into systematic or literature reviews were reviewed but not separately assessed unless additional material not described in the systematic review or narrative literature review was included in the primary study. Relevant data from all primary studies is included in evidence tables. Papers below also included CDR stratified by age and breast density.

In this literature review, two systematic reviews (one including 16 articles generated from five studies), two literature reviews (one including five articles generated from five studies and one review including seven studies) and 12 articles (generated from 12 studies) reported on overall CDR.

Systematic and/or literature reviews

Four reviews: Houssami, (2017); Coop et al., (2016); Hodgson et al., (2016); Houssami, (2015)

RCTs

One study: Maxwell et al., (2017)

Prospective studies

Three articles: Lång et al., (2016A); Bernardi et al., (2016); Caumo et al., (2014)

NB Additional articles reporting on the STORM, STORM-2 and OTS trials are discussed in the systematic and literature reviews.

Retrospective studies (observer performance or single-site analysis)

Eight studies: Pan et al., (2017); Powell et al., (2017); Rodriguez-Ruiz et al., (2017); Conant et al., (2016); McDonald et al. (2016); Sharpe et al., (2016); Wang et al., (2016); Friedewald et al., (2014)

There is strong evidence that CDR is increased when using FFDM + DBT compared to FFDM alone. Increases were reported in a wide range of studies (including large, robust prospective trials) for different combinations of screening strategy including FFDM + DBT, DBT + s2DM, and DBT_{MLO} compared to FFDM alone or DBT_{MLO} + DM_{CC} (the Malmö trial).

CDR results for FFDM + DBT compared to FFDM alone

Systematic reviews

Two systematic reviews explore CDR results for FFDM + DBT compared to FFDM alone (Coop et al., 2016; Hodgson et al., 2016).

The main inclusion criteria for Hodgson et al.'s systematic review were prospective or retrospective studies with more than 1000 women participating in a breast cancer screening program or undergoing opportunistic breast cancer screening. Studies of women with a history of breast cancer or women who were symptomatic or recalled from screening were ineligible. Primary research had to include FFDM as a comparator for DBT alone, FFDM alone compared to FFDM + DBT or FFDM alone compared to DBT + s2DM. Imaging systems used in the primary studies had to be FDA or CE approved. Hodgson et al. included 16 articles generated from five studies (i.e., the OTS and the STORM trials, Friedewald et al.'s 13-site analysis, and two smaller observational studies set in American community radiology practices).

Coop et al.'s systematic review did not describe inclusion criteria in detail, except to note key exclusions (eg, where the comparator was screen mammography, non-English language, publication date pre-2005 which was when DBT became available, etc.). Their systematic review included 21 articles, including papers reporting on the large prospective studies and retrospective analyses reported in Hodgson et al.'s systematic review; however, Coop et al. did not clearly identify which studies were included in the systematic review and which provided contextual information.

Table 8 and Table 9 (pages 36-38) provide a summary of studies that reported on CDR comparing FFDM + DBT to FFDM alone and which were included in either of the systematic reviews. The primary studies included in each of the systematic reviews have considerable differences in study design (for example, small retrospective observational observer performance studies, analyses of screening outcomes or fully paired, prospective trials), sample size, participant characteristics, screening setting (i.e., embedded within a population-based screening program or a single/multi clinic screening setting), screening frequency (i.e., annual or biennial) and/or reading strategy (single/double). These differences in underlying methodology impacted on the pooled analysis undertaken in the systematic reviews. Hodgson et al. only provided pooled analysis of CDR of data from two prospective trials set in population-based screening programs (i.e., data from the STORM and OTS trials). Coop et al. provided no pooled analysis as they considered there to be too much heterogeneity in study design and were limited by the availability of statistically significant results. Neither systematic review provided age stratified data and stratified by breast density although this data is available in the primary literature (and is discussed in *section 3.3* of this report).

Both Coop et al. and Hodgson et al. found that FFDM + DBT increases CDR compared to FFDM alone. In the primary studies included by Hodgson et al. all but one recorded an increase in CDR when using DBT (although not all studies described or achieved statistical significance).

Study Design Quality	Results: overall CDR
Coop et al., 2016 Systematic review Low quality	No pooled analysis
Hodgson et al., 2016 Systematic review Good quality	Difference in CDR: 2.43 per 1000 screening examinations (p<.001; 95%CI: 1.76 to 3.1) Difference in invasive CDR: 2.33 per 1000 screening examinations (p<.001; 95%CI: 1.67 to 3.00)

Hodgson et al.'s pooled analysis of data from the high-quality STORM and OTS trials reported the FFDM + DBT resulted in a difference of 2.43 more cancers detected per 1000 screening examinations (95%CI 1.76-3.1; p<.001). The direction of effect and magnitude are similar for the pooled analysis and the primary data. Incremental increases in CDR with the addition of DBT to FFDM were also reported by Houssami in her 2015 literature review, with a similar upper margin to that calculated by Hodgson et al. Her rates were incremental increases in CDR of between 0.5 to 2.7 cancers per 1000 screening examinations with FFDM + DBT (Houssami, 2015).

Coop et al. and Hodgson et al. reported results from retrospective studies, but no pooled analysis was undertaken (as noted earlier). Reported results suggested the same direction of effect as the pooled analysis from the larger prospective trials. That is, FFDM + DBT detects more cancers than FFDM alone (even if results from retrospective studies were not always statistically significant). In the retrospective observer performance studies, the difference in CDR was smaller than the differences reported in the larger trials. This difference in magnitude of effect may be accounted for by single reading (the European trials employed a double reading strategy but many of the American studies used a single reading strategy), annual instead of biennial screening, bias in the selection of study participants (some samples included women who had a history of breast cancer), studies with sample sizes that were underpowered to detect small differences in CDR or because of potential biases in the allocation of intervention/comparator (eg, women participating in Haas et al.'s study and receiving FFDM alone may have also received DBT during assessment).

Randomised controlled trials

One RCT was identified for this literature review. Maxwell et al. (2017) completed an RCT focused on younger women (aged 40-49 years) with an elevated risk of breast cancer[13] who underwent annual screening for breast cancer. Maxwell et al. reported that FFDM + DBT resulted in increased cancer detection, which is consistent with the results found for other studies; however, the small numbers (1227 women recruited) and annual screening interval may have impacted on the magnitude of these findings. That is, in Maxwell et al.'s RCT, FFDM + DBT detected six cancers compared to five cancers detected with FFDM alone – a difference in detection of one cancer or 4.8 per 1000 screening examinations compared to 4.0 per 1000 screening examinations. This RCT is underpowered to detect small incremental changes to CDR and did not cover a full spectrum of women who would normally be involved in a population-based screening program.

[13] Determined as ≥3% 10-year risk between the ages of 40 and 50 and/or a lifetime risk from age 20 of ≥17%.

Table 11: FFDM + DBT compared to FFDM alone: studies reporting on CDR included in Coop et al. (2016) and Hodgson et al. (2016)

Study	Sample	Study type	**FFDM + DBT** CDR per 1000 screening examinations (95%CI; p-value)	**FFDM alone** CDR per 1000 screening examinations (95%CI; p-value)	**FFDM + DBT** Invasive CDR per 1000 screening examinations (95%CI; p-value)	**FFDM alone** Invasive CDR per 1000 screening examinations (95%CI; p=value)	**Incremental detection** (95%CI)
Prospective trials embedded in European population-based screening programs with biennial screening							
Ciatto et al., 2013 (STORM) *Main article*	7292 asymptomatic, average risk Italian women aged 48 years or older (median age 58 years)	Prospective, fully paired trial using Hologic Selenia Dimensions systems in combination mode	8.1 (6·2–10·4) (p<0.0001 compared to FFDM alone)	5.3 (3·8–7·3)	7.1	4.8	2.7 (1.7-4.2) 53% increase (p=.84)
Caumo et al., 2014 (STORM)		Evaluation of screening metrics at two sites: Trento Verona	7.8 (5.3-10.9) 8.6 (5.6-12.5) (p=.79)	4.9 (3.1-7.5) 5.9 (3.5-9.3) (p=.63)	NR	NR	2.8 (1.5-4.9) 2.6 (1.1-5.2) (p=1)
Houssami et al., 2014 (STORM)		Analysis of STORM data using different reading strategies: Single reading Double reading	7.5 (5.7-9.8; p=.001) 8.1 (6.2-10.4; p=.001)	4.8 (3.3-6.7) 5.3 (3.8-7.3)	NR	NR	2.7 (1.6-4.2) 2.7 (1.6-4.2)
Skaane et al., 2013 (OTS trial)	12,631 Norwegian women aged 50-69 years (average age = 59.3) participating in the biennial Oslo breast screening program, with nine months follow-up.	Prospective, fully paired trial using Hologic Dimensions system with double reading	8.0	6.1	6.4	4.4	40% increase in detection of invasive cancers (p<0.001)
Skaane et al., 2013 (OTS trial)		Prospective, fully paired trial using paired analysis of imaging arms	9.4	7.1	NR	NR	30% increase (p<0.001)
Retrospective, American studies set in community-based radiology practices with annual screening							
Durand et al., 2015	FFDM + DBT: 8591 FFDM: 9364	Retrospective review which includes CAD using Hologic Selenia and Dimensions systems	5.9 (p=.88 compared to FFDM alone) (p=.12 compared to historical control)	5.7 4.4 (historical control)	NR	NR	NR
Lourenco et al., 2015	FFDM + DBT: 12,921 (ages 30.9-89.4 years, average age = 54.6 years) FFDM: 12,577 (ages 29.4-90.6 years, average age = 55.3 years)	Retrospective review of two cohorts (DBT alone=2012/13, FFDM=2011/12), single reading with CAD. FFDM performed using GE Senographe series. DBT performed with Hologic Selenia Dimensions system.	DBT alone 4.6 (p=.44)	5.4 (p=.44)			

Study	Sample	Study type	**FFDM + DBT** CDR per 1000 screening examinations (95%CI; p-value)	**FFDM alone** CDR per 1000 screening examinations (95%CI; p-value)	**FFDM + DBT** Invasive CDR per 1000 screening examinations (95%CI; p-value)	**FFDM alone** Invasive CDR per 1000 screening examinations (95%CI; p=value)	**Incremental detection** (95%CI)
Destounis et al., 2014	524 women aged >30 years (mean age 59 years) including women with a history of breast cancer	Retrospective review of images with double reading. FFDM system was Hologic Selenia or Dimensions, GE Senographe Essential or Fuji CRm. DBT system was Hologic Selenia Dimensions.	5.7	3.8			
Friedewald et al., 2014	FFDM + DBT: 173,663 (ages 52.6-59.7 years, average age = 56.2 years) FFDM: 281,187 (ages 54.4-60.5 years, average age = 57.0 years)	Retrospective review with single reader using data from 13 centres all using Hologic Selenia Dimensions systems	5.4 (4.9 to 6.0; *p*<.001 compared to FFDM alone)	4.2 (3.8 to 4.7)	4.1 (3.7 to 4.5; *p*<.001 compared to FFDM alone)	2.9 (2.5 to 3.2)	28.6% increase
Greenburg et al., 2014	FFDM + DBT: 20,943 FFDM: 38,674 No differences in study arms by age, ethnicity, family history of BC, or prevalence or incidence screening	Retrospective review of mammography outcomes at a multi-site radiology practice using Hologic Selenia or Selenia Dimensions systems	6.3 (*p*=.348)	4.9	4.6 (*p*=.0056)	3.2	28.6% increase (*p*=.035)
McCarthy et al., 2014	FFDM + DBT: 15,571 (average age = 56.7 years) FFDDM: 10,728 (average age = 56.9 years)	Observational study using Hologic Dimensions system for women presenting for mammographic screening at a single institution *NB McDonald et al., (2015) reported on baseline/prevalence CDR based on this study*	5.5 (4.3 to 6.6)	4.6 (3.3 to 5.8)	3.9 (2.9 to 4.8)	3.2 (2.1 to 4.2)	19.6% increase in total CDR (0.9 per 1000 screening examinations (*p*=.32) 21.9% increase in invasive CDR (*p*=.36)
Rose et al., 2013	FFDM + DBT images: 10,878 (average age = 54.5 years) FFDM images: 10,878 (average age = 53.8 years)	Observational reading study of data before/after DBT implemented using Hologic Selenia and Dimensions systems	5.4 (*p*<.0001)	3.5	4.3 (*p*=.07)	2.8	66% increase (*p*<0.0001)
Haas et al., 2013	FFDM + DBT: 6100 FFDM alone: 7058	Retrospective analysis using Hologic Selenia and Dimensions systems *NB: No rate is statistically significant*	Average risk: 5.7 Increased risk: 8.6 Baseline risk: 5.1	Average risk: 5.2 Increased risk: 7.9 Baseline risk: 4.5	NR	NR	NR

[1] 95%CI and p-values listed where noted in the primary papers.
NR = not reported.

Table 12: FFDM + DBT compared to FFDM alone: retrospective observational studies reporting on CDR

Study	Sample	Study type	**FFDM + DBT** CDR per 1000 screening examinations (95%CI; p-value)	**FFDM alone** CDR per 1000 screening examinations (95%CI; p-value)	**FFDM + DBT** Invasive CDR per 1000 screening examinations (95%CI; p-value)	**FFDM alone** Invasive CDR per 1000 screening examinations (95%CI; value)	**Incremental detection** (95%CI)
Pan et al., (2017)	No specific description of the sample provided	Retrospective analysis comparing screening outcomes before/after implementation of DBT (Hologic Dimensions system) from a single hospital site to national data from Taiwan's National Cancer Registry	2012: 8.5 2013: 10.1 2014: 11.4 2015: 8.7	2009: 6.3 2010: 8.1 2011: 7.5	NR	NR	Average of 32.2% increase
Powell et al., 2017	FFDM + DBT: 2304 FFDM: 10,477	Retrospective observational data review of images generated with Hologic's Selenia + Dimensions systems	7.8	5.2	3.5 (1.5 to 6.8) (*p*=.805 compared to FFDM)	3.1 (2.2 to 4.4)	12% difference in invasive CDR
Conant et al., 2016	FFDM + DBT: 55,998 FFDM: 142,883 Women aged 40 to 74 years	Retrospective analysis of data from three PROSPR consortium sites (NB mammography system used not stated)	6.5 (adjusted[2] OR 1.49; 95%CI=1.17 to 1.89; (*p*=.0016)	4.9	4.7 (adjusted[2] OR 1.45; 95%CI=1.09 to 1.92; *p*=.0252)	3.7	34% increase in all cancers, 27% increase in invasive cancer
McDonald et al., 2016	FFDM + DBT: 33,740 FFDM: 10,728 *12079 had one screen* *6293 had two screens* *3023 had three screens*	Retrospective review of mammography metrics from a single site over four years of screening with single reading using Hologic Dimensions system	6.1 (Year 3) 5.8 (Year 2) 5.5 (Year 1) (*p*=.60 compared to FFDM alone)	4.6	NR	NR	34.1% increase
Sharpe et al., 2016	FFDM + DBT: 5703 FFDM: 80,149	Prospective study with a retrospective cohort performed at a single site using Hologic Dimensions system for DBT and GE Senographe Essential, 2000D and DS	5.4 (3.7 to 7.8)	3.5 (3.1 to 3.9)	2.81 (*p*=.61 compared to FFDM)	2.46	54.3% increase (*p*<.0018)
Wang et al., 2016	FFDM + DBT: 12,444 FFDM: 12,444	Retrospective study of DBT and FFDM images (Hologic Dimensions system) of 65 breast cancers	5.2	4.4	3.3	2.6	18% increase in all cancers; increase of 27% for invasive
McDonald et al., 2015	FFDM + DBT (Prevalent): 1859 FFDM + DBT (Incident): 9524 FFDM (Prevalent): 1204 FFDM (Incident): 13,712	Observational study using Hologic Dimensions system for women presenting for mammographic screening at a single institution *NB Data comes from McCarthy et al., 2015*	Prevalent: 5.4 (*p*=.41) Incident: 5.9 (*p*=.51)	Prevalent: 4.6 Incident: 4.2	NR	NR	Prevalent: 40.5% Incident: 17.4% (*p*=.74)

[1] 95%CI and p-values listed where noted in the primary papers.
[2] Adjusted for age, centre, breast density, and prevalent screening examination. NB rate is for exams with at least one year of follow-up.NR = Not reported

Maxwell et al.'s RCT timeframe and annual screening interval could have enabled comparison of CDR at prevalent/incident[14] screening examination and the authors considered undertaking this analysis, but small numbers impacted on the ability to complete this. The authors noted that the higher invasive CDR found in the prospective trials may not be sustained over time due to higher CDR in women participating in screening for the first time (which may have been a large proportion of Maxwell's study population) (also see discussion of McDonald et al.'s 2016 results and results from the Malmö trial results, Lång et al. 2016A).

Prospective screening trials

Most prospective trials comparing FFDM + DBT to FFDM alone were discussed in Hodgson et al. (2016) and Coop et al. (2016).

Study Design Quality	Results: overall CDR
Bernardi et al., 2016 Trial (STORM-2) High quality	FFDM + DBT: 8.5 per 1000 screening examinations (82 cancers detected in 9677 screens) (95%CI: 6.7-10.5) FFDM alone: 6.3 per 1000 screening examinations (p=.0001)

Another recent trial, the STORM-2 trial reported further CDR rates (Bernardi et al., 2016). This single-site Italian fully paired trial involved 9672 women and compared both FFDM + DBT to FFDM alone as well as DBT + s2DM to both FFDM + DBT and DBT alone. The CDR for FFDM + DBT was 8.5 compared to 6.3 for FFDM. The incremental detection rate is 2.2 cancers per 1000 screening examinations (95%CI 1.2-3.3). Further results reporting on CDR for DBT + s2DM are discussed below.

One of the first STORM papers referenced but not discussed by Hodgson et al. (Caumo et al., 2014) noted that CDR results were similar for the two trial centres. FFDM + DBT CDR results from Verona and Trento were 7.8 and 8.6 per 1000 screening examinations compared to FFDM CDR results of 4.9 and 5.9 per 1000 screening examinations. This provides evidence that increased CDR when using FFDM + DBT is likely to be seen across different screening clinic settings. Transferability, however, was not so clear in the Friedewald analysis of data from 13 sites. Friedewald reported considerable variation in CDR between different sites (2.3 to 6.1 per 1000 screening exams for FFDM alone compared to 3.1 to 7.3 per 1000 screening examinations for FFDM + DBT) although an overall increase in CDR was recorded at all sites (Friedewald et al., 2014).

The Malmö trial (Lång et al., 2016A) used a different combination of DBT and digital mammography (a sequential approach to screening: DBT_{MLO} compared to FFDM, plus a reading arm of DBT_{MLO} plus DM_{CC}). CDR results from the Malmö trial are discussed on page 42.

Retrospective observational studies

Since Hodgson et al.'s 2016 systematic review, seven retrospective observational studies have reported CDR for FFDM + DBT compared to FFDM alone. All studies are based in the United States, except one which is based in Taiwan (Pan et al., 2017). None of retrospective observational studies were clearly embedded in a national population-based screening program (that is, most are undertaken in single or multi-site community-based radiology practices or, in

[14] Prevalent screening refers to a woman's first mammogram (also referred to as baseline screening). Incident screening refers to subsequent screening examinations.

Pan et al.'s study, screening outcome data from a single institution was compared to national screening program data).

There is considerable variation in methodology between the retrospective studies included in Hodgson's systematic review. Compared to earlier retrospective studies (which focused on observer performance), more recent studies move towards comparing screening outcomes before and after the implementation of DBT in clinical settings. Other differences include shorter screening interval compared to the prospective studies. Most, but not all, American studies use a one-year screening interval compared to the two-year screening interval in the STORM, OTS, STORM-2 and Malmö trials. The retrospective studies also vary in the description and inclusion of prevalent and incident screening outcome. Inter-observer variability is less likely to be an issue in prospective studies to because of increased familiarity with DBT.

While the included studies were generally large, study samples also varied in size and population characteristics. In one case, the actual study population was unclear (Pan et al., 2017). As with other retrospective studies, limited randomisation was undertaken, and some studies had obvious biases in study participant selection and allocation of study participants to either FFDM alone or FFDM + DBT. For example, Sharpe's comparison of a prospective cohort to retrospective data from women who had received FFDM alone resulted in more women with mammographically dense breasts undergoing FFDM + DBT than in the control group. A noted limitation was that the study sample may be more likely to have cancer than those participants who received FFDM alone. Individual study details are provided in Table 9 (page 38).

Studies published since 2016 tend to have longer periods of follow-up and involve larger groups of asymptomatic women in population-based screening settings compared to the earlier retrospective studies discussed in Coop et al. (2016) and Hodgson et al. (2016). Despite some of the methodological limitations in the later primary studies, all retrospective studies (Pan et al., 2017; Powell et al., 2017; Conant et al., 2016; McDonald et al., 2016; Sharpe et al., 2016; Wang et al., 2016), provided evidence that is consistent with the effect seen in earlier studies and in the prospective trials: FFDM + DBT is a superior test for detecting cancer compared to FFDM alone.

One of the retrospective studies (Conant et al., 2016) reported a statistically significant increase in CDR (adjusted for family history of breast cancer, breast density and prevalent screening exams) of 1.6 cancers per 1000 screening examinations (FFDM + DBT compared to FFDM alone). Other studies reported increases of between 0.8 and 2.6 cancers per 1000 screening examinations (although significance is either not reported or not achieved for these results). Over the studies, FFDM + DBT increased CDR by an average of 34.5 percentage points (range = 27 to 54.3 percentage points). The incremental increase reflects the CDR increases seen in the large prospective trials (i.e., 30-40% increases reported in Lång et al., 2016A, Ciatto et al., 2013 and Skaane et al., 2013).

McDonald et al. (2016) completed an analysis of changes in CDR over a three-year period. The authors reported a non-significant increase in CDR from 4.6 per 1000 screening examinations to 5.5, 5.8 and 6.1 per 1000 screening examinations at FFDM + DBT Years 1, 2, and 3 with a corresponding rise in PPV_1. They reported an analysis of OR for CDR over three years noting non-significant fluctuations in CDR between Year 1 and Year 3 (i.e., in Year 1: FFDM compared to FFDM + DBT: OR=1.35 [0.93-1.94]; Year 2: OR=1.28 [0.88-1.85]; Year 3: OR=1.35 [0.93-1.94]; overall p=.80). McDonald et al.'s (2016) study also enabled an initial exploration of the impact of prevalent and incident screening on CDR by screening strategy. They found that CDR decreased from 13.2 per 1000 screening examinations for women with one screen to 6.2 per 1000 screening examinations for women who underwent two examinations. This may reflect the

impact of lower CDR for incident screening examinations compared to prevalent examinations. The lower Year 2 result, as noted by McDonald et al., is still higher than the CDR for FFDM alone (4.9 per 1000 screening examinations).

Related to this, an earlier retrospective study reported individual outcomes following implementation of DBT (McCarthy et al., 2014). Using data from McCarthy et al.'s study, McDonald et al. (2015) found a higher incremental CDR for FFDM + DBT compared to FFDM alone for both prevalent and incident screens (although the results did not achieve significance); however, the improvement in incremental detection for prevalent screening was 40.5% compared to 17.5% for women undergoing incident screening. Given these results, the authors concluded that FFDM + DBT may be best targeted towards women undergoing their first screening because of the higher incremental CDR seen in this group. Further discussion about DBT's impact on women by age band is discussing in *section 3.3* of this report.

Pan et al.'s 2017 study recorded CDR for four years post-implementation of DBT, finding that the CDR fluctuated but generally rose between the first and last year of study; however, possible reasons for this are not explored in their analysis.

CDR results for DBT + s2DM compared to FFDM + DBT and FFDM alone

Systematic reviews, RCTs and literature reviews

No systematic reviews or RCTs compared CDR from DBT + s2DM compared to FFDM + DBT or FFDM alone.

One literature review (Houssami, 2017) summarised the CDR results of studies that investigated a screening strategy based on DBT + s2DM compared to FFDM alone. In August 2017, Houssami conducted a literature review of population-based breast cancer screening-related clinical outcomes (including CDR) for the following imaging combinations: DBT + s2DM compared to either FFDM + DBT or FFDM alone. The literature review included papers reporting on the STORM-2 and OTS trials (Bernardi et al., 2016; Skaane et al., 2014) and three retrospective studies (Aujero et al., 2017; Freer et al., 2017; Zuckerman et al., 2016). Data from five other retrospective studies was also presented but not discussed in detail by Houssami as these studies were smaller (four studies had samples of <400), were based on enriched data sets and/or were not specifically focused on screening (i.e., these studies provided evidence to inform a diagnostic or assessment setting and did not discuss CDR). *Allen + Clarke's* search (with a publication close date of 31 December 2017) did not identify any further published evidence investigating DBT + s2DM and cancer detection that was not published in Houssami. As such, we present the overall data on CDR as discussed in Houssami's paper.

CDR results presented in Houssami's literature review are described in Table 10 (see page 43).

All study results reported that DBT + s2DM resulted in superior CDR compared to FFDM alone. Comparable CDR results were achieved for DBT + s2DM compared to FFDM + DBT, although the findings were mixed. Some studies (including the STORM-2 trial) reported a small increase in CDR for DBT + s2DM compared to FFDM + DBT. Statistically significant CDR results from the STORM-2 trial shows that DBT + s2DM performs better than FFDM alone (8.8 cancers detected per 1000 screening examinations compared to 6.3 respectively; $p<.0001$) (Bernardi et al., 2016). Compared to FFDM + DBT, Bernardi et al. reported a non-significant increase of 0.3 favouring DBT + s2DM ($p=.58$). Data from the OTS trial also demonstrated comparable (but not statistically significant) performance between DBT + s2DM and FFDM + DBT (*NB* these results

are from the "after" arm, that is following the implementation of improved processing software).

The three retrospective studies discussed in Houssami's literature review all reported that FFDM + DBT had a higher CDR compared to DBT + s2DM (although all these results are non-significant). It is important that these results are further replicated in other studies.

There is strong evidence that FFDM + DBT has higher CDR results compared to FFDM alone. Emerging CDR results are comparable for FFDM + DBT and DBT +s2DM but much of the data on CDR for DBT + s2DM did not achieve statistical significance. Despite this, Houssami concludes that, because there is limited difference in the CDR between DBT + s2DM and FFDM + DBT (and that CDR are higher in the DBT strategies compared to mammography alone) coupled with a lowered radiation dose when using s2DM (see *section 3.4*), breast screening programs should consider integrating s2DM.

DBT + s2DM looks like a promising approach for increased cancer detection; however, further robust evidence that demonstrate that DBT + s2DM's comparability or superiority to FFDM + DBT in terms of cancer detection is needed. This may come from upcoming large studies like the PROSPECTS and TOSYMA trials and the Italian RCTs.

CDR results for DBT_{MLO} compared to other imaging combinations

Two prospective studies investigated DBT_{MLO} to other combinations of digital mammography and DBT: Rodriguez-Ruiz et al., (2017) and Lång et al., (2016A).

The active Malmö trial is using a different combination of screening strategies compared to many of the other studies included in this literature review. Instead of FFDM + DBT or DBT + s2DM, Lång et al. used a sequential approach to screening: DBT_{MLO} compared to FFDM, plus a reading arm of DBT_{MLO} plus DM_{CC}. They are also using a different DBT system (Siemens Mammomat Inspirations). An explorative analysis and interim CDR results are available from the Malmö trial, a fully paired prospective trial embedded in Sweden's national screening program (Lång et al., 2016A). Results are available for the first half of study participants (7500 women).

Study Design Quality	Results: CDR
Lång et al., 2016A Trial (Malmö) High quality	CDR: DBT_{MLO} + DM_{CC}: 8.9 (95%CI 6.9-11.3; p<.0001) FFDM: 6.3 (95%CI 4.6-8.3; p<.0001) PPV: 24% for FFDM and DBT 21 cancers detected in the DBT arm alone. One was detected in the FFDM arm alone.

Although the screening strategy and DBT system used in this trial differs from other studies, interim results from the Malmö trial confirm the direction of effect for CDR seen in other studies. The interim results reported are a statistically significant 43 percentage point increase in cancer detection compared to FFDM alone (p<.0001). Lång et al. noted that DBT_{MLO} alone (without FFDM or DM_{CC}) detected every cancer that was also detected in the two-view arm. They concluded that, from these interim results, it may be possible to use DBT alone without digital mammography for screening purposes in asymptomatic women; however, we note that other issues such as recall rate and radiation dose must be considered (and balanced) before such a step is sensible.

Rodriguez-Ruiz et al. (2017) compared the performance of DBT_{MLO} to three other screening strategies (FFDM alone, DBT_{MLO} + DM_{CC}, and FFDM + DBT). While the case set included both asymptomatic and symptomatic women (and CDR was not the reported outcome), the authors reported that sensitivity of DBT_{MLO} was not inferior to the other screening strategies (72%

Table 13: Studies investigating DBT + s2DM included in Houssami's 2017 literature review

Study	Sample	Study type	**DBT + s2DM** CDR per 1000 screening examinations (95%CI; p-value)	**FFDM + DBT** CDR per 1000 screening examinations (95%CI; p-value)	**FFDM alone** Invasive CDR per 1000 screening examinations (95%CI; p-value)	**DBT + s2DM** Invasive CDR per 1000 screening examinations (95%CI; p=value)	**FFDM + DBT / FFDM alone** Invasive CDR per 1000 screening examinations (95%CI; p=value)
Prospective trials embedded in European population-based screening programs with biennial screening							
Bernardi et al., 2016 (STORM-2)	9672 asymptomatic Italian women aged 49 years or older (median age 58 years) who attended population-based screening	Prospective, fully paired trial using Selenia Dimensions system with C-view for FFDM + DBT with single reading (*NB analysis of 9677 women*)	8.8 (7.0 to 10.8; *p*<.0001 compared to FFDM; (*p*=.58 compared to FFDM + DBT)	8.5 (6.7 to 10.5; *p*<.0001 compared to FFDM)	6.3 (4.8 to 8.1)	NR	NR
Skaane et al., 2014 (OTS trial)	12,270 screens from 24,901 Norwegian women aged 50-69 years (mean age 59.2 years)	Prospective, fully paired trial using Selenia Dimensions system with C-view for FFDM + DBT with double reading and two study periods (one using C-view, one using an earlier software. Only rates using C-view are reported here)	Study period 1: 7.4 Study period 2: 7.8 Study period 1 = decrease of 7% Study period 2 = decrease of 2%	Study period 1: 8.0 Study period 2: 7.7	NR	NR	NR
Retrospective American studies set in community-based radiology practices with annual screening							
Aujero et al., 2017	Mammograms from a single USA practice: 16,173 mammograms with DBT + s2DM; 30,561 mammograms with FFDM + DBT; 32,076 mammograms with FFDM alone	Retrospective observational study with single reading using Selenia Dimensions system with C-view	6.1 (*p*=.27 compared to FFDM; *p*=.71 compared to FFDM + DBT)	6.4 (OR, 1.21; 95%CI: 0.98 to 1.48)	5.3	76.5% (*p*=<.01 compared to FFDM + DBT)	FFDM + DBT: 61.3%
Freer et al., 2017	31,979 women receiving a screening mammogram a single USA practice between 10/2013–12/2015 (9525 women screened with DBT + s2DM; 1019 screened with FFDM + DBT; 21,435 screened with FFDM alone	Retrospective analysis using Hologic Selenia and Dimensions systems with C-view	5.9 (non-adjusted) 5.4 (adjusted)[3]	6.9 (non-adjusted) 5.7 (adjusted)[3]	5.9 (non-adjusted) 5.0 (adjusted)[3] (For adjusted CDR: *p*=.66 compared to FFDM; *p*=.90 compared to FFDM + DBT	4.6 (non-adjusted) 4.3 (adjusted)[3]	*Non-adjusted* FFDM + DBT: 3.9 FFDM alone: 4.3 *Adjusted*[3] FFDM + DBT: 3.4 FFDM alone: 3.9
Zuckerman et al., 2016	FFDM + DBT: 15,571 DBT + s2DM: 5366	Observational study set in a community screening setting using Hologic Dimensions system	5.03 (*p*=.72 compared to FFDM + DBT)	5.45	NA	3.85	4.10 (*p*=.84 compared to FFDM + DBT)

[1] PPV, 95%CI and p-values listed where noted in the primary papers.
[2] PPV_1 = positive predictive value for cancer cases per recalled patient; PPV_2 = positive predictive value for biopsy recommended; PPV_3 = positive predictive value for biopsy perform
[3] Adjusted controlling for priors, age, density, and effect of reader
NR = not reported

compared to 76% for FFDM alone and FFDM + DBT). There was no statistical difference in jackknife free-response receiver operating characteristic (JAFROC) curves between DBT_{MLO} and the other screening strategies.

3.1.1. Invasive CDR and other characteristics

This literature review describes invasive CDR findings from 12 articles (from eight studies) including pooled analysis from one systematic review and three literature reviews/systematic reviews. Primary studies already incorporated into systematic or literature reviews were reviewed but not separately assessed unless additional material not described in the systematic or literature review was included in the primary study. Relevant data from all primary studies is included in Tables 8 and 9 (see pages 36-38). Papers below also include information stratifying data by age and breast density. Stratification is discussed further in *section 3.3*.

Systematic and/or literature reviews

Three reviews: Skaane (2017); Coop et al., (2016); Hodgson et al., (2016)

RCTs

None identified.

Prospective studies

Three studies: Bernardi and Houssami., (2017); Bernardi et al., (2016); Lång et al., (2016A)

NB Additional articles reporting on the STORM, STORM-2 and OTS trials are also discussed in the systematic and literature reviews.

Retrospective studies (observer performance or single-site analysis)

Six studies: Aujero et al., (2017); Freer et al., (2017); Powell et al., (2017); Conant et al., (2016); Sharpe et al., (2016); Wang et al., (2016)

There is evidence that invasive CDR is increased when using DBT compared to FFDM alone. Increases were reported in a range of studies (including large, robust prospective trials) for different combinations of screening strategies including FFDM + DBT, DBT + s2DM, and DBT_{MLO} compared to FFDM alone (the Malmö trial). Houssami et al. (2016) noted that incremental increases in CDR are predominantly due to increased invasive CDR, which reflects the findings of this literature review.

Invasive CDR results for FFDM + DBT compared to FFDM alone

Systematic reviews

Both Coop et al. (2016) and Hodgson et al. (2016) reported that FFDM + DBT resulted in an increase in invasive CDR. Hodgson et al.'s (2016) fixed effect meta-analysis of data from the STORM and OTS trials found a statistically significant increase in invasive CDR detected by FFDM + DBT compared to FFDM alone. This resulted in the detection of an additional 2.33 cancers per 1000 screening examinations (95%CI 1.76-3.1; p<.001). This is only slightly lower than the overall additional cancers per 1000 screening examinations reported in this systematic review (2.43 cancers per 1000 screening examinations).

Hodgson et al. also reported data from retrospective studies, most of which only looked at CDR rather than results for specific types of cancer (including invasive CDR). Hodgson et al. reported results consistent (if smaller) with the prospective trials. All studies reported increases in invasive CDR ranging from 0.7 to 1.5 additional invasive cancers detected per 1000 screening examinations (FFDM + DBT compared to FFDM) (Friedewald et al., 2014; Greenburg et al., 2014; McCarthy et al., 2014; Rose et al., 2013). Friedewald's multi-site analysis was the only study to achieve statistical significance for invasive CDR. That study reported an invasive CDR of 4.1 per 1000 screening examinations compared to 2.9 for FFDM alone (p=.001). Statistical significance was not achieved for invasive CDR results from other retrospective studies. Also 95%CI were quite wide but overall the results reported an increase in invasive CDR favouring FFDM + DBT. Possible reasons for the lower invasive CDR in retrospective studies (compared to the results seen in the OTS and STORM trials) are discussed in *section 3.1*.

Hodgson et al. (2016) found a slight (but not significant) increase in detection for non-invasive cancers with the FFDM + DBT compared to FFDM alone. Coop et al. (2016) noted that FFDM has a higher sensitivity for the detection of DCIS microcalcifications compared to DBT and concluded that DBT does not increase detection of DCIS. This may be because the microcalcifications are more easily seen in 2D and may be more difficult to see in the multiple 1mm slices created with DBT imaging. Coop et al. do not distinguish DCIS by lower or higher histological grade.

Coop et al. (2016) also reported invasive CDR from smaller and older retrospective studies (although the review was largely silent on invasive CDR). The authors noted that increased invasive CDR for soft tissue cancers presenting without microcalcifications were found in two studies (although the increases noted in the primary studies did not achieve significance). This led the authors to conclude that DBT appears to support superior detection of invasive cancers that do not present with microcalcifications.

Literature describing DBT's performance in relation to microcalcifications is discussed in the following sub-section, *Other cancer characteristics.*

Retrospective studies

All retrospective studies reporting on invasive CDR reported that invasive CDR increased when DBT is used as an adjunct to FFDM (i.e., FFDM + DBT). This included studies based on outcomes analysis from clinical practices as well as observer performance (Powell et al., 2017; Conant et al., 2016; Sharpe et al., 2016; Wang et al., 2016). Tables 8 and 9 (see pages 36-38) provide detail on each study. Increases ranged from 0.35 to 1 additional cancer per 1000 screening examinations. Only Conant et al.'s result (1 additional cancer per 1000 screens) from their retrospective analysis of data from three PROSPR consortium sites achieved statistical significance. While the reported rates were smaller than those reported for the prospective trials (covered in Hodgson et al.'s systematic review), taken together, the evidence base suggests that use of DBT as an adjunct screen increases detection of invasive cancer compared to FFDM alone.

Invasive CDR results for DBT + s2DM compared to FFDM alone

Prospective studies

Results from the STORM-2 trial (Bernardi et al., 2016) reported that FFDM + DBT results in an increase in invasive CDR. The STORM-2 trial detected 74 invasive cancers out of 90 total cancers; however, no invasive CDR was calculated (and it is not clear whether the increase in invasive cancers detected related to FFDM + DBT or DBT + s2DM). Results from STORM-2 were also

reported for mean tumour size for invasive cancers. Mean tumour size for invasive cancers detected with all three screening strategies (FFDM alone, FFDM + DBT and DBT + s2DM) were similar (12.7 mm, standard deviation 7.8). Invasive cancers detected with only FFDM + DBT or DBT + s2DM were slightly smaller: 11.6 mm (SD 9.4).

Bernardi and Houssami (2017) described cancers detected in the STORM-2 trial. They reported that DBT detected most small invasive cancers (depicted as irregular masses or distortions) that could not be seen on FFDM images (although the authors noted that these were also difficult to detect with DBT). Interesting, they also noted that one of the cases detected was invasive lobular cancer, a cancer that is hard to detect in any screening modality.

Retrospective studies

Freer et al.'s 2017 study compared single reading of DBT + s2DM images to FFDM and FFDM + DBT and reported non-significant findings for invasive CDR of 4.3, 3.4 and 3.9 per 1000 screening examinations for DBT + s2DM, FFDM + DBT and FFDM respectively. This finding was not repeated in other studies. Aujero et al. (2017) also reported on invasive CDR, finding that DBT + s2DM detected 4.64 cancers per 1000 screening examinations compared to 3.21 cancers per 1000 screening examinations when FFDM alone was used (76.5% of invasive cancers compared to 61.3% detected by FFDM + DBT; p=.01, OR 2.06, 95%CI 1.19-3.56). The PPV for DBT + s2DM was also significantly higher than FFDM + DBT (40.8% compared to 28.5%; p=.0001; OR 1.15, 95%CI 0.90-48). This suggests that DBT + s2DM may be better able to detect invasive cancers compared to FFDM + DBT; however, the results from the two studies are not consistent.

Invasive CDR results for DBT_{MLO} compared to other imaging combinations

Invasive CDR is not reported in the Malmö interim results; however, data tables show that the DBT total arm (DBT_{MLO+CC} + prior digital mammography) detected 17 more invasive cancers compared to FFDM alone arm (58 compared to 41 cancers). Lång et al. (2016A) noted that there were no statistically significant differences between those cancers detected with DBT_{MLO} or by the FFDM reading arm. The PPV was the same for both reading arms, which differs from other study findings (which have tended to report a difference that favours DBT).

Other cancer characteristics

Studies focused on diagnostic populations were excluded from this literature review; however, much of the literature discussing tumour characteristics is based on either diagnostic populations or datasets enriched with cancer cases. This sub-section provides a short summary based on other narrative literature reviews which report on cancer characteristics. A more fulsome exploration of the relationship between different combinations of DBT and digital mammography imaging will be provided in a companion literature review on the role of DBT in the assessment and diagnosis of breast cancer (to be prepared by *Allen + Clarke* in mid-2018).

Two literature reviews summarised findings on other cancer characteristics (Skaane, 2017; Houssami et al., 2016). The literature reported varying results about tumour characteristics. Some studies discussed by Houssami et al. (2016) suggested that there are differences in the cancers detected with DBT compared to those detected with FFDM alone. DBT appears to detect smaller cancers which are more likely to be invasive, and cited data from STORM, Malmö and the OTS trial noting that DBT appears to detect microcalcifications as well as FFDM but that incremental detection focuses on invasive cancers. Other studies, including data from the OTS

trial, reported no differences in grade, size or radiologic signs for cancers detected either with DBT or FFDM but the OTS trial did report that the incremental cancers detected with FFDM + DBT were predominantly invasive.

Additionally, in Houssami et al.'s pictorial atlas, the authors described the cancers detected with DBT (based on a literature review and data from the STORM, OTS, Malmö and the ASTOUND trials) (Houssami et al., 2016). They reported that DBT improves conspicuity (due to reduced tissue overlap) for certain types of lesions including spiculated or stellate masses and architectural distortions compared to those seen with FFDM alone. This improved conspicuity likely underpins the increased detection of invasive cancers (as reported above). Houssami et al. also noted that benign findings (including radial scars) were easier to detect with DBT; however, improved conspicuity may result in additional assessment of suspicious areas that are benign (particularly stellate distortions), which may increase false positive rates overall. This was an issue reported particularly with the screening strategy used in the Malmö trial (DBT_{MLO} compared to FFDM).

Detection of clinically significant microcalcifications appears to be an area of unsettled science. Skaane (2017) reported that some older studies indicated FFDM's superior performance in terms of detection of microcalcifications compared to DBT. Citing Kopans et al.'s (2011) study, which found equal or greater clarity of microcalcifications, Skaane (2017) reported that classification of these radiologic features might have clinical significance; however, Houssami et al. (2016) argued that trials have reported comparable detection of microcalcifications for both DBT and FFDM.

Tumour size at detection was also explored by Houssami et al. (2016). Reporting on data from the OTS and Malmö trials, the authors noted that DBT may detect more smaller sized and lower grade cancers compared to FFDM alone. Radial scar may also be more visible on DBT compared to FFDM, which could result in additional unnecessary assessments to determine malignancy. Skaane (2017) concluded that over-diagnosis may be an issue of perception, requiring improved treatment decision tools rather than reduced initial detection (eg, over-treatment of abnormalities that may never be clinically significant is also an issue rather than over-diagnosis alone). This is an area that requires additional research to unpick the relationship between more timely diagnosis of asymptomatic cancers and the risk of overdiagnosis.

3.1.2. Interval cancer rates

The interval cancer rate refers to the number of breast cancers that become symptomatic within 12 months of a mammogram in which no abnormalities were detected. Improved cancer detection (either by detecting invasive cancers earlier or improved lesion conspicuity) can lead to a decline in interval cancers. Higher interval cancer rates may be due to visibility issues associated with FFDM, interpretation failure (i.e., images are not read correctly), interpretation protocols or reading strategy (i.e., double reading reduces the risk of an abnormality being missed by up to 15% (Houssami et al., 2017), or they may reflect a cancer that is new and was not previously visible on mammogram. The interval cancer rate can be used as a surrogate indicator for screening benefit and is a marker of the sensitivity of a breast screening program.

Both Coop et al. (2016) and Hodgson et al. (2016) noted that there is limited information about the impact of DBT (either alone or integrated with FFDM) on interval cancer rate within population-based screening settings. Insufficient data on interval rates was included in the

primary studies informing the systematic reviews so no pooled analysis was completed by either Hodgson et al. or Coop et al.

Three studies (including one RCT and one prospective, fully paired trial) reported on interval cancer rate: Maxwell et al. (2017); Houssami et al., (2017) and McDonald et al. (2016). Overall, there is insufficient evidence to determine what impact DBT (either as an adjunct to FFDM or with s2DM) may have on reducing the interval cancer rate. No studies reported on interval cancer rate for DBT alone.

Randomised controlled trials

While interval cancer rate was not a primary outcome measure for this RCT, Maxwell et al. (2017) recorded two cases of interval breast cancer following examination with FFDM + DBT in high-risk younger women aged 40-49 years. The authors did not provide data for FFDM alone nor do they present data as "CDR per 1000 screening examinations". This is a small RCT (1227 women) and was underpowered to detect small incremental differences in interval cancer rate. Maxwell et al.'s findings relating to interval cancer do not add significantly to the body of knowledge about the impact of FFDM + DBT on longer term mortality outcomes at this time.

Study Design Quality	Results: interval cancer rate
Maxwell et al., 2017 RCT Low quality (SIGN)	Total cancers detected: 11 Interval cancers: 2 following FFDM + DBT examination *NB ICR of 'per 1000 screening examinations' not used. No rate for FFDM alone is presented.*

Prospective screening trials

The STORM-2 trial reported data on interval cancers (Houssami et al., 2014). Based on 13 months follow-up, the authors reported the overall first year interval cancer rate was six cancers or 0.82 per 1000 screening examinations (95%CI 0.30–1.79). It is not clear whether this is an overall rate or whether it relates specifically to a reading strategy (single or double) or to the use of adjunct DBT. Further evidence on interval cancers from STORM-2 will be published in 2018.

Retrospective observational studies

One retrospective observational study (McDonald et al., 2016) provided data on interval cancer rates. McDonald et al. used state cancer registry data to calculate the interval cancer rate between one screening examination when FFDM was used and the first year of FFDM + DBT screening. They reported that FFDM + DBT resulted in a lower interval cancer rate (0.5 per 1000 screening examinations with FFDM + DBT compared to 0.7 per 1000 screening examinations with FFDM alone); however, this finding is not statistically significant. We also know that some of the same women participated in up to three of the four cohorts (3,023 participants in total). The authors note that the limited change in interval cancer rate means that FFDM + DBT is detecting clinically relevant cancers.

Study Design Quality	Results: interval cancer rate
McDonald et al., 2016 Retrospective review Adequate quality (SIGN)	0.7 per 1000 women screened with FFDM alone in Year 0; 0.5 per 1000 women screened with FFDM + DBT in Year 1 (*p=.60*; 95%CI not provided)

Although there is some evidence to suggest that FFDM + DBT reduces interval cancer compared to FFDM alone, early results must be further replicated in other longer-term studies such as the Bergen trial, the Tomosynthesis Mammographic Imaging Screening Trial (Canada) and the two proposed Italian RCTs (all described in *section 1.4* of this report).

3.1.3. Sensitivity

Because follow-up and overall study timeframes in the primary studies tends to be shorter, few estimates of absolute sensitivity for FFDM + DBT are available. Instead, five articles reported on the *relative* sensitivity of FFDM + DBT compared to FFDM alone (four articles reported data from the STORM trial and one reported from PROSPR consortium). One article discussed the *relative* sensitivity of DBT + s2DM compared to FFDM and one study looked at DBT_{MLO}.

Relative sensitivity of FFDM + DBT compared to FFDM alone

Mixed results are reported on the *relative* sensitivity of FFDM + DBT compared to FFDM alone.

One prospective trial (STORM) found an increase in *relative* sensitivity; however, a large retrospective analysis from PROSPR consortium sites did not.

One article based on the STORM trial data (Ciatto et al., 2013) provided an indication of FFDM + DBT sensitivity within a breast screening program setting. In this study, double reading of FFDM + DBT resulted in an increased sensitivity compared to FFDM alone: FFDM + DBT: 90.77% (95%CI 80.7-96.51) compared to 60.00% for FFDM alone (95%CI 41.10-71.96%). These results were replicated in Houssami et al.'s 2014 research, which also reported on a single reading strategy (which has a lower overall sensitivity compared to double reading), but which still reported improved sensitivity for FFDM + DBT compared to FFDM alone. The impact of incidence and prevalence screening is not considered in these studies (although the potential impact of that effect is noted).

Study Design Quality	Results: *relative* sensitivity of FFDM + DBT compared to FFDM alone
Bernardi et al., 2014 Trial (STORM) High quality (SIGN)	FFDM + DBT: 87% FFDM alone: 63%
Houssami et al., 2014 Trial (STORM) High quality (SIGN)	*Single reading* FFDM + DBT: 85% (74-92) FFDM alone: 54% (41-66) *Double reading* FFDM + DBT: 91% (81-97) FFDM alone: 60% (47-72)
Ciatto et al., 2013 Trial (STORM) High quality (SIGN)	FFDM + DBT: 90.77% FFDM alone: 60.00%
Conant et al., 2016 Retrospective analysis	FFDM + DBT: 90.9% FFDM alone: 90.6%

Bernardi et al. (2014) and Bernardi et al. (2012) reported further STORM trial data on the relative sensitivity of FFDM + DBT compared to FFDM alone based on individual reader performance. Bernardi et al. (2014) investigated the performance of eight radiologists who interpreted 14,525 screens. For FFDM alone, the number of cancers detected from total cancers in the sample varied substantially by individual radiologist (38% to 83%; median = 63%). The range was much narrower (and had a higher median) for FFDM + DBT (78% to 93%; median = 87%). Although individual performance varied, all but one radiologists' performance improved

when using FFDM + DBT compared to FFDM alone (one radiologist recorded 83% for both screening strategies). Bernardi et al. (2014) went on to conclude that the greatest improvements were seen in radiologists with the lowest sensitivity at FFDM alone. In an earlier paper reporting on STORM trial data, Bernardi et al. (2012) also investigated three radiologists' performance and found increased incremental detection of 20.8% with FFDM + DBT. Considering overall sensitivity and inter-reader variability, results from the STORM trial support FFDM + DBT being a more sensitive test that detects more cancers compared to FFDM alone and that it enables these to be more easily detected by readers. Further discussion about reader performance is included in *section 4.2*.

Conant et al.'s 2016 retrospective analysis of mammogram images from 198,881 women from three sites in PROSPR consortium noted that sensitivity was not improved between FFDM + DBT (90.9%) compared to FFDM alone (90.6%) (adjusted OR=0.79, 95%CI 0.38–1.64). Reported sensitivity for FFDM alone is much higher than the rate reported in the STORM trial.

Relative sensitivity of DBT + s2DM compared to FFDM + DBT and FFDM alone

One study reported on the *relative* sensitivity of DBT + s2DM compared to FFDM + DBT. Gur et al. (2012) completed a small retrospective observer performance study using an enriched dataset to compare the performance of 10 radiologists when interpreting 114 images created by DBT + s2DM or FFDM + DBT. For fixed reader effect, Gur et al. reported a statistically significant finding that FFDM + DBT was superior in terms of sensitivity compared to DBT + s2DM by 5.4%. The sensitivity outcome included both pathologically proven cancers (n=48) as well as high-risk lesions (n=6) being recalled. When high-risk lesions were removed, FFDM + DBT still had a higher sensitivity (4% more compared to DBT + s2DM, *p*=.05). Evidence from Gur et al.'s study predates the development of Hologic's C-view image reconstruction software and, as such, the study findings may no longer be relevant.

Study Design Quality	Results: *relative* sensitivity of DBT + s2DM
Gur et al., 2012 Retrospective observer performance study	FFDM + DBT: 82.6% DBT + s2DM: 77.2% Difference for fixed reader effect: 5.4% (*p*=.017); for random reader effect (*p*=.053)

Relative sensitivity for DBT_{MLO} compared to other imaging combinations

Our literature review returned one study investigating the relative sensitivity of DBT_{MLO} alone compared to DBT_{MLO} + DM_{CC}, FFDM + DBT, or FFDM alone (Rodriguez-Ruiz et al., 2017). This was a small retrospective study using an enriched sample of 181 women either recalled from screening (33%) or presenting with a clinically significant symptom (67%). We included this study because its findings have an application to the screening context. Rodriguez-Ruiz et al.'s study is also performed on Siemens Mammomat system (not Hologic). Rodriguez-Ruiz et al. found, over six readers, that there was limited difference in sensitivity between the screening strategies. DBT_{MLO} had a sensitivity of 72% (95%CI 68-76) compared to sensitivity of 75% or 76% (95%CI 72-80) for the other screening strategies. Given the small size and methodological design of this single study, it is not possible to determine whether single-view DBT is an effective screening test. Further research is needed.

3.1.4. Summary

There is strong evidence that CDR increases when using DBT compared to FFDM alone. Increases were reported in a range of studies (including large prospective trials) for different combinations of screening strategy including FFDM + DBT, DBT + s2DM, and DBT_{MLO} compared to DM_{CC} or FFDM alone. The direction of effect is consistent across study design, setting and location although there is some variance in magnitude of effect.

here was very limited data about the long-term mortality benefits, treatment morbidity or quality of life improvements associated with FFDM + DBT as a screening strategy. Almost no data exists on results for incident screening compared to prevalent screening, mortality benefit or surrogate indictors of this. Reliable data on interval cancer rate is also scarce.

3.2. Specificity

Specificity (the proportion of people correctly identified as not having breast cancer, or the true negative/positive rate) is an important dimension of an effective population-based breast screening program. The BSA's National Accreditation Standard requires recall rates of less than 10 percent for prevalent screening and less than five percent for incident screening.

A positive initial or final assessment of a suspected cancer is a true positive if it is followed by a biopsy that confirms breast cancer. It is a false positive if no breast cancer is diagnosed within a specified follow-up time (usually about 12 months). A negative initial or final assessment is a false negative if a breast cancer is subsequently diagnosed within a specified follow-up time (and is often indicated through interval cancer rates for a screening program – see *section 3.1.3* for a discussion of interval cancer rates). It is a true negative if it is not (i.e., cancer is not detected within that time). We want to know, based on current evidence, what role DBT plays in a modern breast cancer screening environment and which screening strategy (DBT alone or integrated with other FFDM imaging) is best able to reduce false-positives and unnecessary recalls from screening for women who do not have breast cancer.

Key findings

All retrospective studies show that both FFDM + DBT and DBT + s2DM reduce overall recall rates and false positive recall rates compared to FFDM alone. However, larger prospective study results have reported inconsistent results, with some reporting increased recall rates with the addition of DBT. This is set against a backdrop of generally low recall rates in programs where the trials are embedded (where perhaps we may not expect to see a further decline in rate). Most of the studies investigating recall rates have short timeframes ($\leq$24 months) and recall rates appear to be affected by reading strategy and arbitration mechanisms (which could account for the differences in results). Further research is needed to assess the impact that having previous images available to use in s2DM and DBT interpretation has on recall rates as this may also support an overall decrease in rates. Over time, it is likely that the overall and false positive recall rates associated with FFDM + DBT and DBT + s2DM will reduce as the readers become more familiar with the images and potentially different display of parenchymal features.

3.2.1. Overall recall rates

The overall recall rate is the percentage of women asked to return for follow-up assessment after an abnormality is detected during screening (i.e., any recall resulting in true or false

positive findings). This literature review describes overall recall rate findings from 14 articles (generated from eight studies) including pooled analysis from two systematic reviews and two literature reviews.

Systematic and/or literature reviews

Four reviews: Houssami, (2017); Coop et al., (2016); Hodgson et al., (2016); Vedantham et al., 2015.

RCTs

One study: Maxwell et al., (2017)

Prospective studies

Five studies: Bernardi et al., (2016); Lång et al., (2016A); Lång et al., (2016B); Dang et al., (2014); Skaane et al., (2014)

NB Additional articles reporting on the STORM and STORM-2 and OTS trials are also discussed in the systematic and literature reviews.

Retrospective studies (observer performance or single-site analysis)

Four studies: Rodriguez-Ruiz et al., (2017); Shin et al., (2015); Sumkin et al., (2015); Gur et al., (2012).

Primary studies already incorporated into systematic or literature reviews were reviewed but not separately assessed unless additional material not described in the systematic or literature review was included in the primary study. Relevant data from all primary studies is included in evidence tables.

The literature is not settled about the association between DBT and recall rates. Two main complexities exist:

1. The impact of reading strategy on recall rates (rather than issues with image acquisition or quality): some studies report increased recall rates with FFDM + DBT; others report reductions, and different studies use different interpretation or arbitration protocols which may influence recall rates, and
2. The impact on recall rates when used in a high-quality screening program that is already achieving a low recall rate.

Interpretation timings and other implementation issues are discussed in Chapter 4.

Recall rate results for FFDM + DBT compared to FFDM alone

Systematic reviews

Two systematic reviews explored recall rates for FFDM + DBT compared to FFDM alone (Coop et al., 2016; Hodgson et al., 2016). The main inclusion criteria used by Coop et al. and Hodgson et al. and comments on study design strengths and limitations of the studies used in the systematic reviews are described in *section 3.1.1*. For overall recall rates, both systematic reviews included only a small range of studies (i.e., the STORM and OTS trials, which both systematic reviews report on, and some of the larger observational studies investigating DBT's role in screening). Table 11 (page 54) provides a summary of included studies for the two systematic reviews.

Neither systematic review includes pooled analysis for recall rate (i.e., they provide narrative coverage only). An attempt to perform meta-analysis on the recall rate data by Hodgson et al. resulted in significant heterogeneity (I^2 = 89%): a summary effect was not calculated. Neither systematic review provided age stratified recall rate data or data stratified by breast density although this data is available in the primary studies (and is discussed in *section 3.3* of this report). Both Hodgson et al. and Coop et al.'s conclusions are based on studies with short timeframes: robust data on mortality benefit (or surrogate indicators of this) is not available at this time. Further research with pooled analysis would be useful to better understand the impact of DBT and reading strategy on specificity.

Overall recall rates reported in the systematic reviews

Both Coop et al. and Hodgson et al. found that results from the retrospective US studies consistently showed that FFDM + DBT has a significantly lower overall recall rate compared to FFDM alone; however, the results from European prospective trials were mixed. Differences in findings between the STORM and OTS trials may be because recall rates varied according to the double reading strategy adopted or because the overall recall rates in these screening programs are already low. Results from the STORM trial are consistent with those from the retrospective studies, and pre-arbitration results from the OTS trial reflect what might be found in a single reader program.

Data from the STORM trial (double sequential reading with FFDM first followed by FFDM + DBT) reported that FFDM + DBT resulted in a statistically significant reduction of 0.7% in overall recall rate compared to FFDM alone. The STORM trial recalled women if either radiologist reported a positive finding.

The OTS trial reported recall results by reading arm and pre- or post-implementation process. The OTS trial conducted an arbitration meeting for the FFDM + DBT and FFDM alone arms of the study. Pre-arbitration overall recall rates of individual readers (like that found in a single reader strategy) were higher for FFDM + DBT (2.78% for FFDM + DBT compared with 2.1% for FFDM alone). Post-arbitration (i.e., more like a double reading strategy), higher recall rates than pre-arbitration were observed for FFDM + DBT compared to FFDM (3.67% for FFDM + DBT compared with 2.9% for FFDM alone as reported in both Skaane et al. 2013A and 2013B). Post-arbitration, this translates to FFDM + DBT having 6.2 more recalls per 1000 screens than FFDM alone. Hodgson et al. reported that DBT images were available at the arbitration meeting for both the FFDM and FFDM + DBT arms, which biases the results in favour of a lower recall rate with FFDM as radiologists had more information to work with and suggests that the recall rate in the FFDM arm may have been underestimated (i.e., the readers had more information with which to decide).

Both Coop et al. and Hodgson et al. reported results from retrospective studies published before 2015, which all described a similar direction of effect as the results from the STORM trial (that is, adjunct screening with FFDM + DBT resulted in statistically significant lower recall rates than FFDM alone). These retrospective studies tended to use a single reading strategy. Lourenco et al. (2015) reported a very significant reduction in overall recall rates with DBT of 31% (p<.00001) and noted that this was consistent with previous findings from Rose et al. (2013), Skaane et al. (2013) and Haas et al. (2013). Friedewald et al. (2014) reported a lower (but still statistically significant) reduction in recall rates for FFDM + DBT compared to FFDM alone (16.1%, p<.001). The other retrospective studies had consistent direction of effect although the magnitude of

effect was higher than that reported by STORM (from 15% reduction in recall rates to 40% reduction). This may reflect the impact of single reading.

Studies reporting on overall recall rates as reviewed by Coop et al. and Hodgson et al. are summarised in Table 11 (overleaf). Recall rates are presented as the percentage of women that were recalled out of the total number of screens for each cohort.

Table 11: Studies reporting on overall recall rate included in Coop et al. (2016) and Hodgson et al. (2016)

Study	Sample	Study type	FFDM + DBT Recall rate	FFDM alone Recall rate	Difference between recall rates
Prospective trials embedded in European population-based screening programs with biennial screening					
Ciatto et al., 2013 (STORM)	7292 asymptomatic, average risk women	Prospective, fully paired trial	4.3%	5.0%	NR
Skaane et al., 2013 (OTS trial) (single reading)	12,631 women aged 50-69 years participating in the biennial Oslo breast screening program, with nine months follow-up.	Prospective, fully paired trial using Selenia Dimensions unit with double reading	2.78%	2.1%	NR
Skaane et al., 2013 (OTS trial) (double reading)		Prospective, fully paired trial using Selenia Dimensions unit + paired analysis of imaging arms	3.67%	2.9%	NR
Retrospective, American studies set in community-based radiology practices with annual screening					
Lourenco et al., 2015	12,577 FFDM and 12,921 DBT exams	Retrospective review of two cohorts (DBT alone=2012/13, FFDM=2011/12, single reading)	6.4%	9.3%	31% decrease (p<.00001)
Destounis et al., 2014	524 women aged >30 years (mean age 59 years)	Retrospective review of images with double reading	4.2% (p<0.0001)	11.45%	NR
Friedewald et al., 2014	173,663 images from 13 different sites	Retrospective review with single reader	9.1%	10.7%	16.1% decrease (p<.001)
Rose et al., 2014	10,878 FFDM+DBT images matched to 10,878 FFDM images	Observational reading study of data before/after DBT implemented	5.5%	8.7%	-
Haas et al., 2013	13,158 women at one of four clinical sites	Retrospective analysis *NB: No statistical significance*	8.4%	12%	29.7% decrease (p<.01)

Randomised controlled trials

The RCT completed by Maxwell et al. (2017) investigated recall results for women aged 40-49 years, and is therefore analysed with respect to age, below, at *section 3.3*.

Prospective studies

Dang et al. (2014) conducted a prospective trial focused on the real-world clinical performance of implementing DBT into a screening program. This study used a single reading strategy. The differences in interpretation times between FFDM and FFDM + DBT for multiple participating radiologists was quantified. A total of 3665 routine screening examinations performed during a six-month timeframe were interpreted in at least five sessions per radiologist per screening strategy. Although it was not the primary focus, the authors also noted that there were significantly increased recall rates for FFDM compared with FFDM + DBT (6.3% and 5.3% respectively; $p<.0001$). This equates to an approximate incremental decrease of 15% in favour of the DBT screening strategy.

Sumkin et al. (2015) reported much higher recall rates for both FFDM and FFDM + DBT compared with other studies. A single reading strategy was used. The authors noted the high recall rate of prevalent screening examinations for both the FFDM + DBT and FFDM alone screening strategies as a limitation of the study. It is likely that higher recalls would be seen in a screening population with a high rate of prevalent screening examinations because previous (mammographic or DBT) images are not available. This mean less data for radiologists to use to detect benign features that may look suspicious. Sumkin's study also had limited generalisability due to the study being conducted at one institution with a single group of radiologists on a specific group of women ($N=1080$). The authors reported a statistically significant reduction in recall rates between FFDM and FFDM + DBT of 33% ($p<.001$).

Bernardi et al. (2016), reporting on results from the STORM-2 trial, did not report specific overall recall rates for FFDM + DBT compared with FFDM. See *section 3.2.2* for their results on false positive recall rates.

Retrospective observational studies

Retrospective observational studies published since Coop et al. and Hodgson et al.'s 2016 systematic reviews all favoured a positive impact on recall rates with the introduction of DBT (and these results align to the STORM trial). They all used single reading approaches.

Powell et al. (2017) reported on a retrospective observational study using databases to compare overall rates of recall for 12,781 women (10,477 undergoing FFDM and 2304 undergoing FFDM + DBT). Powell et al. found that the addition of DBT to FFDM resulted in significantly lower recall rates. FFDM + DBT had an overall recall rate of 14%, and the FFDM only group had a recall rate of 16%, reflecting a 12.5% reduction in the overall recall rate with the addition of DBT ($p=.017$). This means that for women screened with FFDM + DBT, 14% of women were recalled for further screening. This was a non-randomised study with relatively small sample sizes. Therefore, although reduced recall rates were statistically significant, the study lacked statistical power overall. Durand et al. (2015) similarly reported a statistically significant reduction in recall rates between FFDM alone and FFDM + DBT (36.6% reduction; $p<.0001$), for women recalled for either asymmetries or calcifications. A single reading strategy was used.

Retrospective studies reporting on overall recall rates are summarised in Table 12 (overleaf). Recall rates are reported as the percentage of women who were recalled from each cohort.

Table 12: Retrospective observational studies reporting on overall recall rate

Study	**Sample**	**Study type**	**FFDM + DBT** Recall rate (95%CI; p-value)	**FFDM alone** Recall rate (95%CI; p-value)	**Difference between recall rates** (95%CI; p-value)
Pan et al. 2017	No specific description of the sample provided	Retrospective analysis comparing screening outcomes before/after implementation of DBT (Hologic Dimensions system) from a single hospital site to national data from Taiwan's National Cancer Registry	9.0%-10.1%	11.4% - 12.2%	17.8% decrease (p<.01)
Powell et al. 2017	FFDM + DBT: 2304 FFDM: 10,477	Retrospective observational data review of images generated with Hologic's Selenia + Dimensions systems: single reading	14%	16%	12.5% decrease (p=.017)
Conant et al., 2016	FFDM + DBT: 55,998 FFDM: 142,883	Retrospective analysis of data from three PROSPR consortium sites (NB mammography system used not stated): single reading	8.7%	10.4%	15.6% decrease (p<.0001)
Durand et al., 2015	8591 FFDM + DBT mammograms and 9364 FFDM alone mammograms (n=17,955)	Retrospective review: single reading	7.8%	12.3%	36.6% decrease (p<.0001)
McDonald et al. 2015	FFDM + DBT (Prevalent): 1859 FFDM + DBT (Incident): 9524 FFDM (Prevalent): 1204 FFDM (Incident): 13,712	Observational study using Hologic Dimensions system for women presenting for mammographic screening at a single institution: single reading	8.8%	10.4%	22% decrease (p=.002)
Greenburg et al., 2014	38,674 FFDM examinations; 20,943 FFDM + DBT	Retrospective review of mammography outcomes at a multi-site radiology practice	13.6%	16.2%	13.6% decrease (p<.0001)
McCarthy et al., 2014	15,571 women screened with FFDM + DBT and 10,728 screened with FFDM	Single read observational study using Hologic Dimensions system for women presenting for mammographic screening at a single institution	8.8%	10.4%	16% decrease (p<.001)

Results for DBT + s2DM compared to FFDM + DBT and FFDM alone

Three studies (all discussed in Houssami's 2017 literature review) reported recall rates for DBT + s2DM compared to FFDM + DBT and FFDM alone. Houssami did not perform a pooled analysis of overall recall rates. The method of the review is described in *section 3.1.1*.

In her narrative review, statistically significant overall recall rates from three retrospective American studies (Aujero et al., 2017; Freer et al. 2017; Zuckerman et al. 2016) showed a range of different results were reported depending on screening strategy comparator.

Two studies (Aujero et., 2017; Freer et al., 2017) showed recall rates for FFDM alone being higher than both FFDM + FFDM and DBT + s2DM.

- Aujero et al. reported recall rates of 8.7% for FFDM alone, 5.8% for FFDM + DBT, and 4.3% for DBT + s2DM. These results were statistically significant.
- Freer et al. reported a statistically significant reduction in recall rate between FFDM (7.8%) and FFDM + DBT (6.39%; p<.001). However, while there was a decrease in recall rates between FFDM + DBT and DBT + s2DM (5.52%), this result was not statistically significant (*p*=.25).

Houssami concluded that the significantly lower recall rates reported in Aujero et al. using DBT + s2DM compared with FFDM + DBT were probably reflective of improved interpretation of DBT, given that the readers had gained experience in FFDM + DBT before transitioning to DBT + s2DM.

Zuckerman et al. (2016) reported only on the difference between FFDM + DBT and DBT + s2DM and found a statistically significant decrease in recall rates between the two screening strategies (8.8% for FFDM + DBT compared to 7.1% for DBT + s2DM; *p*<.001).

Bernardi et al. (2016), reporting on results from the STORM-2 trial, did not report specific overall recall rates for DBT + s2DM compared to FFDM + DBT and FFDM alone (see *section 3.2.2* for their results on false positive recall rates).

Table 13 (below) summarises the overall recall rates as reported in Houssami's literature review.

Table 13: Studies included in Houssami's 2017 literature review

Study	Sample	Study type	**DBT + s2DM** Recall rates (95%CI; *p*-value)	**FFDM + DBT** Recall rates (95%CI; *p*-value)	**FFDM alone** Recall rates (95%CI; *p*-value)
Retrospective American studies set in community-based radiology practices with annual screening					
Aujero et al., 2017	Mammograms from a single USA practice: 16,173 mammograms with DBT + s2DM; 30,561 mammograms with FFDM + DBT; 32,076 mammograms with FFDM alone	Retrospective observational study with single reading using Selenia Dimensions system with C-view	4.3% (statistically significant decrease vs. FFDM + DBT and FFDM: *p*<.001)	5.8% (statistically significant decrease vs. FFDM: *p*<.001)	8.7%
Freer et al., 2017	31,979 women receiving a screening mammogram a single USA practice between 10/2013–12/2015 (9525 women screened with DBT + s2DM; 1019 screened with FFDM + DBT; 21,435 screened with FFDM alone	Retrospective analysis using Hologic Selenia and Dimensions systems with C-view	5.52% (decrease vs. FFDM + DBT non-significant: *p*=.25)	6.39% (statistically significant decrease vs. FFDM: *p*<.001)	7.83%
Zuckerman et al., 2016	15,571 American women screened with FFDM + DBT and 5366 women screened with DBT + s2DM	Observational study set in a community screening setting using Hologic Dimensions system	7.1% (statistically significant decrease vs.	8.8%	NR

Study	Sample	Study type	DBT + s2DM Recall rates (95%CI; *p*-value)	FFDM + DBT Recall rates (95%CI; *p*-value)	FFDM alone Recall rates (95%CI; *p*-value)
			FFDM + DBT: *p*<.001)		

Recall rates results for DBT_{MLO} compared to other imaging combinations

Study Design Quality	Results: recall rates
Lång et al., 2016A Trial High quality	Overall recall rates: DBT_{MLO} + DM_{CC}: 3.8% (3.3 to 4.2) FFDM: 2.6% (2.3 to 3.0) Increase in recall rate using DBT relative to DM: 43% (26 to 62; *p*<.0001)

Interim results from the Malmö trial (Lång et al., 2016A) are interesting. Using a sequential approach to screening (DBT_{MLO} compared to FFDM, plus a reading arm of DBT_{MLO} plus DM_{CC}), Lång et al., (2016A) reported a statistically significant increase in overall recall rates. The recall rate for the DBT reading arm was 3.8% (95%CI 3.3-4.2) compared with 2.6% (95%CI 2.3-3.0) for FFDM alone. The increase in recall rate when using DBT compared with FFDM was 43% (95%CI 26-52; *p*<.0001). Lång et al. reported that DBT appeared to be particularly sensitive to the detection of small spiculated lesions (discussed in *section 3.1.2*). These tended to be either low-grade cancers or benign radial scars. This, combined with the findings on CDR (discussed in *section 3.1.1*), led the authors to conclude that DBT will enhance detection of benign lesions and sometimes areas of normal breast parenchyma. Lång et al., thought that this probably contributed to the significant increase in recall rate reported with DBT_{MLO} compared to FFDM. It is also important to note that this screening program has a low recall rate in general anyway. The recall rate observed in this study was higher than estimated for the sample size calculation. Lång et al. (2016A) did not consider that this increase was due to the screening strategies used. It would be reasonable to expect that the increase in overall recall rate would be lower in the second half of the study population, due to increased experience of the readers. This will be reported on in 2018.

3.2.2. False positive rate

False positive mammogram results are concerning for both women and breast screening program administrators. Women who are recalled for further investigation may experience high levels of anxiety, along with the inconvenience and expense of attending a further appointment which may bring no health benefit to the woman or may lead to her undergoing additional and unneeded invasive tests and/or biopsy. The health system may incur unnecessary costs based on biopsy and further assessment of suspected abnormalities which turn out to be benign.

False positive recalls occur particularly in younger women aged under 50 years and those with other risk factors for breast cancer such as high breast density. One of the reasons for this is that younger women generally have more dense breasts than older women. This might result in more overlap of glandular tissue which can produce composite densities which may appear like cancers. Further discussion on the effect of age and breast density on screening outcomes is included in *section 3.3*.

False positive rate results for FFDM + DBT compared to FFDM alone

Seven studies reported false positive rates for FFDM + DBT compared to FFDM alone.

Systematic reviews

One systematic review (Hodgson et al., 2016) explored false positive recall rates for FFDM + DBT compared to FFDM alone. The main inclusion criteria used by Hodgson et al. and comments on study design strengths and limitations of the studies used in the systematic reviews are described in *section 3.1.1*.

Studies included in Hodgson et al. all reported a lower false positive rate for FFDM + DBT compared to FFDM alone. The review identified two prospective trials (OTS and STORM) and three retrospective studies conducted in the US (Destounis et al., 2014; Friedewald et al., 2014; Lourenco et al., 2015). Hodgson et al. did not perform meta-analysis on either the European or US studies in terms of false positive rates (for the reasons previously noted in *section 3.1.1*).

As with overall recall rates, Hodgson et al. reported that STORM and OTS trials observed different results for false positive rate when using FFDM + DBT compared to FFDM alone.

- In STORM, lower recall rates and lower false positive rates were observed when using FFDM + DBT compared to FFDM alone. FFDM + DBT had 9.1 less false positives per 1000 screens compared to FFDM (95%CI: -11.8 to -7.2).
- In the OTS trial, lower false positive rates using FFDM + DBT were found pre-arbitration (difference per 1000 screens was -8 for FFDM + DBT vs. FFDM alone), but higher false positive rates found post-arbitration (and higher recall rates were observed overall). After consensus by arbitration, the difference for FFDM + DBT versus FFDM was +5.4 per 1000 screens for false positives (95%CI: 4.2-6.8).

No pooled analysis for overall false positive rates was provided because attempts to combine the results using meta-analysis results in significant heterogeneity ($I^2 = 99\%$).

The reported results of the retrospective US studies described a similar direction of effect as the results from the STORM trial (that is, FFDM + DBT results in lower false positive rates than FFDM alone). However, the magnitude of the difference in false positives varied so drastically that no meta-analysis was performed on this group either. Destounis et al. (2014) reported a highly statistically significant decrease in false recalls between FFDM and FFDM + DBT of 74.74 per 1000 screening examinations (95%CI 105.6 to -43.1). Lourenco et al. (2015) and Friedewald et al. (2014) reported less extreme, although still statistically significant reductions when using FFDM + DBT compared to FFDM alone: 28.7 false recalls per 1000 screens (95%CI -35.1 to -22.2) and 17.4 per 1000 screens (95%CI 15.6-19.2) respectively.

Table 14 (overleaf) provides a summary of the false positive recall rates reported in Hodgson et al. Where percentages are given, this indicates the percentage of overall screens that resulted in a false positive recall. False positive recalls were reported as percentages in some studies, and false positive screens per 1000 in others. The measurement for different results is specified.

Randomised controlled trials

The RCT completed by Maxwell et al. (2017) reported results for younger women only, and is therefore analysed with respect to age, below, at *section 3.3.3*.

Table 14: Studies reporting on false positive rate included in Hodgson et al. (2016)

Study	Sample	Study type	**FFDM + DBT** FPR (95%CI; p-value)	**FFDM alone** FPR (95%CI; p-value)	**Difference between FPR** (95%CI; p-value)
Prospective trials embedded in European population-based screening programs with biennial screening					
Ciatto et al., 2013 (STORM)	7292 asymptomatic, average risk women	Prospective, fully paired trial	3.5%	4.4%	9.3 decrease per 1000 screens (-11.8 to -7.2)
Skaane et al., 2013 (OTS trial)	12,631 women aged 50 -69 years participating in the biennial Oslo breast screening program, with nine months follow-up.	Prospective, fully paired trial using Selenia Dimensions unit with double reading	8.5%	10.3%	Pre-arbitration: 8 decrease per 1000 screens Post-arbitration: 5.4 increase per 1000 screens (4.2 to 6.8)
Retrospective American studies set in community-based radiology practices with annual screening					
Lourenco et al., 2015	12,577 FFDM and 12,921 DBT examinations	Retrospective review of two cohorts (DBT alone=2012/13, FFDM=2011/12), single reading	5.94%	8.80%	28.7 decrease per 1000 screens (-35.1 to -22.2)
Friedewald et al., 2014	173,663 images from 13 different radiology sites	Retrospective review with single reader	8.4%	10.14%	17.4 decrease per 1000 screens (-15.6 to -19.2)
Destounis et al., 2014	524 women aged >30 years (mean age 59 years)	Retrospective review of images with double reading	3.63%	11.07%	74.4 decrease per 1000 screens (-105.6 to -43.1)
Rose et al., 2014	10,878 FFDM+DBT images and 10,878 matched FFDM images	Observational reading study of data before/after DBT implemented	NR	NR	29.5 decrease per 1000 screens

Retrospective observational studies

Since Hodgson et al.'s 2016 systematic review, no relevant retrospective observational studies have reported on overall false positive recall rates for FFDM + DBT. Therefore, the false positive recall rates included in this review have been noted above.

Results for DBT + s2DM compared to FFDM + DBT and FFDM alone

Houssami (2017) reported on false positive recall rates for DBT + s2DM compared to FFDM + DBT or FFDM alone (see Table 15, overleaf). The method of the review is described in *section 3.1.1*. Statistically significant false positive recall rates from one retrospective American study (Aujero et al., 2017) showed that false positive recall rates are significantly reduced when using DBT + FFDM compared to FFDM alone (5.2% compared to 8.2%; OR 0.61, 95%CI 0.58-0.66). DBT + s2DM had the lowest false positive recall rates: 3.6% (OR 0.69, 95%CI 0.62-0.76 compared to DBT + FFDM).

The results from the OTS trial (Skaane et al., 2014) reported a minimal, statistically insignificant decrease in false positive recall between FFDM + DBT and DBT + s2DM (that is, a false positive rate of 4.5% when using DBT + s2DM compared to 4.6% when using FFDM + DBT; p=.85).

However, the data from Bernardi et al. (2016) showed a false positive recall rate for DBT + s2DM that was significantly greater than those for FFDM + DBT and FFDM alone. The recall rate for FFDM alone was 3.42%, FFDM + DBT was 3.97% (p<.001 vs FFDM), and DBT + s2DM was 4.45% (p<.001 vs FFDM and p=.03 vs FFDM + DBT). Houssami noted that this data should be considered in context. The difference in rates was small (about -0.5%), and it is likely that it resulted from incorporating s2DM into real-world screening practice for the first time without previous experience with s2DM images relative to the established experiences with FFDM in the screening program involved in the trial.

Vedantham et al. (2015) conducted a literature review which generally corroborates the results from Houssami. Reporting on Skaane et al. (2014), Vedantham et al. noted that with an early version of the s2DM algorithm used in period 1 of the STORM trial, DBT + s2DM statistically differed from FFDM + DBT in false positive recall rate, whereas a newer version applied in period 2 of the STORM trial, the false positive rate was not statistically different between DBT + s2DM and FFDM + DBT. This shows progressive improvement in algorithms used for generating s2DM images.

Table 15: Studies included in Houssami's 2017 literature review

Study	Sample	Study type	DBT + s2DM False positive rate (95%CI; p-value)	FFDM + DBT False positive rate (95%CI; p-value)	FFDM alone False positive rate (95%CI; p-value)
Prospective trials embedded in European population-based screening programs with biennial screening					
Bernardi et al., 2016 (STORM-2)	9672 asymptomatic Italian women aged 49 years or older (median age 58 years) who attended population-based screening	Prospective, fully paired trial using Selenia Dimensions system with C-view for FFDM + DBT with single reading	4.45% (statistically significant increase vs. FFDM: *p*<.001; and FFDM + DBT: *p*=.03)	3.97% (statistically significant increase vs. FFDM: *p*<.001)	3.42%
Skaane et al., 2014 (OTS trial)	12,270 screens from 24,901 Norwegian women aged 50-69 years (mean age 59.2 years)	Prospective, fully paired trial using Selenia Dimensions system with C-view for FFDM + DBT with double reading and two study periods (one using C-view, one using an earlier software. Only rates using C-view are reported here)	4.5% (non-significant increase vs. FFDM + DBT: *p*=.85)	4.6%	NR for sub-analysis
Retrospective American studies set in community-based radiology practices with annual screening					
Aujero et al., 2017	Mammograms from a single USA practice: 16,173 mammograms with DBT + s2DM; 30,561 mammograms with FFDM + DBT; 32,076 mammograms with FFDM alone	Retrospective observational study with single reading using Selenia Dimensions system with C-view	3.6% (statistically significant decrease vs. FFDM + DBT and FFDM: *p*<.001)	5.2% (statistically significant decrease vs. FFDM: *p*<.001)	8.2%

Prospective screening trials: secondary analysis from STORM-2

All false positive recall rates

A secondary analysis from the STORM-2 trial indicated that false positive recall rates for FFDM + DBT or s2DM + FFDM are significantly lower than false positive rates for FFDM (Houssami et al., 2016). The false positive rates reported were as follows:

- FFDM alone: 3.42% (95%CI 3.07-3.80)
- FFDM + DBT: 2.60% (95%CI 2.29-2.94) and
- DBT + s2DM: 2.76% (95%CI 2.45-3.11).

FFDM + DBT had a false positive recall rate of 0.82% less than FFDM (95%CI -1.17 to -0.48; *p*<.001) and DBT + S2DM had a false positive recall rate of 0.66% less than FFDM (95%CI -1.07 to -0.25; *p*=.002). However, the false positive recall rate for DBT + s2DM was still slightly higher than that for FFDM + DBT. The authors did not provide a reason for this result. These results indicate that Houssami's contextual analysis of the initial STORM (above) results on false positive recall was probably right, and that further experience using s2DM relative to FFDM has resulted in reductions in false positive rates ().

Retrospective observational studies

Gur et al. conducted a study in 2012 to retrospectively compare the performance of synthetically reconstructed 2D images in combination with DBT versus FFDM + DBT. This study reported effectively the same false positive recall rate for FFDM + DBT (29.8%) compared with DBT + s2DM (29.7%). The authors noted that a limitation of the study was the lack of prior examinations available for viewing and interpretations. This was a very early utilisation of s2DM, which may account for the comparatively higher false positive recall rates, when looking at other studies.

False positive recall rates for DBT_{MLO} compared to other imaging combinations

Interim results from the Malmö trial, (Lång et al., 2016B) reported a statistically significant decrease in false positive recall rate when using DBT_{MLO} alone compared with FFDM + prior mammogram images (when available) during the first 1.5 years of the trial. The false positive recall rate for DBT_{MLO} alone after arbitration was 1.7%, and for digital mammography alone was 0.9%. For women recalled on both FFDM and DBT_{MLO}, the false positive recall rate was 1.1%. Importantly, the average false positive recall rate for DBT alone over the 1.5 years of the trial (reported to date) was 1.9% (1.5-3.3), for digital mammography alone it was 0.9% (0.4-1.2) and for FFDM + DBT was 1.0% (0.6-1.5).

Study Design Quality	Results: false positive rate
Lång et al., 2016 Trial High quality	All false positive rates: DBT_{MLO} alone: 1.7% DM alone: 0.9% FFDM + DBT: 1.1%

The authors noted that the over-time reduction (to stabilise at 1.5%) in false positive recall rates implied that specificity can be improved with increased reader experience in interpreting DBT images. All false positive rates reported in this study were post-arbitration since that reflects the actual impact on clinical practice. Overall, these false positive recall rates are very low.

3.2.3. Specificity

This review identified eight articles that reported on the *relative* specificity of FFDM + DBT compared to FFDM alone. Most study results demonstrate an improvement in specificity with the addition of DBT. One article reported on the relative specificity of DBT + s2DM compared to FFDM.

Hodgson et al. (2016) reported an overall increase in specificity from the STORM trial (Ciatto et al., 2013) of FFDM + DBT (96.49% specificity; 95%CI 96.04-96.90) compared with FFDM alone (95.55% specificity; 95%CI 95.04-96.01). Poplack et al. (2017) reported a mean specificity of 88.9% for FFDM + DBT, and 84.8% for FFDM alone (CI were not reported for these results).

The mean specificities for the three readers in Shin et al. (2015) were not significantly different between FFDM and FFDM + DBT, although the results did show that FFDM + DBT had a slightly lower specificity than FFDM alone (78.5% for FFDM; 75.1% for FFDM + DBT; p=.260). Two readers showed a decrease in specificity with the combined technique, and one reader showed an increase. The authors noted that these results were different from those reported in other studies and stated that the overall recall rate decreased with the use of DBT. The lack of improvement in specificity was explained by the fact that the number of benign or normal cases was small, and that the data were enriched with malignant lesions.

Results from Rodriguez-Ruiz et al. (2017) showed no significant difference between FFDM + DBT and FFDM for specificity (p=.553). It was noted that for the group of DBT experienced radiologists, no statistically significant difference in specificity was observed (p=.482). For the inexperienced group, specificity was slightly higher for FFDM + DBT compared to FFDM, but again, the result was not significant (p=.777).

Lång et al. (2016B) commented that the drop in false positive recall rates over the first 1.5 years indicated that the specificity can be improved with increased reader experience but did not provide any specific specificity rates. Haas et al. (2013) commented that the addition of DBT resulted in improvements to specificity, without observing large increases in sensitivity. Gennaro et al. (2013) reported that specificity with FFDM + DBT was higher (84.9%) than FFDM alone (83.0%); however, the difference was not statistically significant (0.018; 95%CI -0.008-0.044; p=.130). Conant et al. (2016) reported a statistically significant increase in specificity between FFDM + DBT and FFDM (91.3% for FFDM + DBT; 89.7% for FFDM for a 1.39% difference; 95%CI: 1.30-1.48; p<.0001).

Gur et al. (2012) reported the same specificity levels when interpreting FFDM + DBT compared to DBT + s2DM. It was noted that improved s2DM images would possibly be available soon.

3.2.4. Positive predictive value (PPV)

PPV is the probability that asymptomatic women with a screening mammogram that detects something suspicious (i.e., a "positive" mammogram) are subsequently diagnosed with breast cancer. It is a measure of the overall accuracy of the screening test. The three dimensions of PPV reported in a breast screening context are described in the box (right).

PPV_1: number of verified attributable cancers per number of women recalled from screening

PPV_2: the number of cancers diagnosed per the number of biopsies recommended

PPV_3: the number of cancers diagnosed per the number of biopsies performed

Primary studies included in Hodgson et al. and Coop et al.'s systematic reviews report information on PPV; however, the systematic review authors only note that PPV for FFDM + DBT was generally comparable to or improved in the larger prospective trials compared to FFDM (that is, they do not discuss the PPV results specifically). As such, we have reviewed the primary research to report on PPV in more detail. In this literature review, 17 studies reported on PPV.

Systematic and/or literature reviews

None of the systematic reviews discuss PPV in detail.

RCTs

No RCTs discussed PPV.

Prospective studies

Two studies: Lång et al. (2016A); Skaane et al. (2013B)

Retrospective studies (observer performance or single-site analysis)

15 studies: Aujero et al. (2017); Freer et al. (2017); Pan et al. (2017); Powell et al. (2017); Rafferty et al. (2017); Conant et al. (2016); McDonald et al. (2016); Zuckerman et al. (2016); Lourenco et al. (2015); McDonald et al., (2015); Destounis et al. (2014); Friedewald et al. (2014); Greenburg et al. (2014); McCarthy et al. (2014); Skaane et al. (2014)

Some studies did not report on all three dimensions of PPV (see data in Table 16, Table 17 and Table 18, pages 67-70). Information about the prevalence of breast cancer in the study participant populations was not provided in any article. In addition, studies generally reported limited information about the number of women (or mammogram images) separated into prevalent or incident screening. Only one study (McDonald et al., 2015) reported results but these did not achieve significance. Limited information regarding interval cancer rate is available in the included studies (see *section 3.1.3*). As such, we compare PPV between the different screening strategies (FFDM + DBT compared to FFDM alone, or DBT + s2DM compared to FFDM + DBT or FFDM alone), rather than as absolute indicators of DBT's sensitivity or specificity as a screening test.

None of the prospective trials reported statistically significant $PPV_{1\text{-}3}$.

PPV_1 (cancers diagnosed per the number of women recalled from screening)

Eleven retrospective studies reported on PPV_1 for FFDM + DBT compared to FFDM alone.

In all studies, PPV_1 was higher for FFDM + DBT compared to FFDM alone although there is variation in the magnitude of effect. Statistically significant results for PPV_1 were noted in seven studies with results ranging 3.4 to 6.7% for FFDM + DBT compared to 3.0 to 4.4% for FFDM alone. The overall mean increase between FFDM + DBT and FFDM alone across the six studies that achieved significance was 2.1%. Pan et al. (2017) report an increase of 4% in favour of FFDM + DBT; however, the statistical significance of their findings is not reported.

Rafferty et al.'s large study of screening performance metrics from 454,322 women (13 centres in the PROSPR consortium) investigated PPV_1 by age band. They report that FFDM + DBT increased PPV_1 for women in all age groups, although the increase is at least 1.1% higher for women aged over 50 years compared to younger women. This difference in effects continues to increase as women age and is highest for women aged 60-69 years.

Four studies report on PPV_1 comparing FFDM + DBT to DBT + s2DM (Aujero et al., 2017; Freer et al., 2017; Zuckerman et al., 2016; Skaane et al., 2014). Results show that compared to FFDM + DBT, DBT + s2DM increases the number of cancers detected per recalled women with the range of increase varying from 0.9 to 3.4 percent. Aujero et al. (2017) also reported that the PPV for DBT + s2DM was significantly higher than FFDM + DBT (40.8% compared to 28.5%; p=.0001; OR 1.15, 95%CI 0.90-48). Skaane et al. (2014) also demonstrated an increase in PPV_1 over time, which may reflect improvements in reader performance when using DBT for screening. Because DBT + s2DM out-performed FFDM + DBT in terms of PPV_1 and given that FFDM + DBT had higher PPV_1 rates compared with FFDM alone, it is reasonable to assume that DBT + s2DM would also show an increase the number of cancers detected per number of recalls compared to FFDM alone. This result is demonstrated in Aujero et al.'s study, which observes a statistically significant increase of 8.3%.

Overall results on PPV_1 indicate that, on average, FFDM + DBT accurately detected proportionally more women recalled from screening who had breast cancer compared to FFDM

alone. Six studies with significant findings reported that on average, recalls based on FFDM+DBT screening are correctly identifying an additional two women with diagnosable breast cancer for every 100 women recalled, compared with FFDM alone.

Additionally, DBT + s2DM shows promise of having further increased accuracy compared to FFDM alone. One study (Aujero et al., 2017) found that recalls based on FFDM + DBT screening are correctly identifying an additional eight women with diagnosable breast cancer for every 100 women recalled, compared with FFDM alone. Recall rates for DBT + s2DM were 4.3%. The size of this effect indicates that DBT + s2DM may be more accurate than either FFDM or FFDM + DBT in identifying the need for recall than FFDM + DBT. This suggestion is supported by a small number of studies (Freer et al., 2017; Zuckerman et al., 2016; Skaane et al., 2014) indicating that recalls based on DBT + s2DM screening are correctly identifying between one and three more women with diagnosable breast cancer for every 100 women recalled, compared with recalls based on FFDM + DBT screening.

PPV_2 and PPV_3 (cancers diagnosed per the number of biopsies recommended or biopsies performed, respectively)

Eleven retrospective studies reported on PPV_2 (*N*=3) and/or PPV_3 (*N*=9) for FFDM + DBT compared to FFDM alone. The three studies reporting on PPV_2 (Pan et al., 2017; Powell et al., 2017; McCarthy et al., 2014) all showed an increase in PPV_2 for FFDM + DBT compared to FFDM alone; however, the statistical significance of the differences was either not reported (Pan et al., 2017; McCarthy et al., 2016) or did not achieve significance (Powell et al., 2017). This indicates that biopsies recommended by FFDM + DBT screening are potentially more likely to result in a diagnosis of breast cancer than biopsies recommended by FFDM screening alone; however, lack of information about statistical significance is an issue.

Only Friedewald et al. (2014) reported statistically significant results for PPV_3, finding that FFDM had a lower PPV_3 compared to FFDM + DBT. This result is consistent (in terms of direction of effect) with the other studies that reported non-significant increases in PPV_3 for FFDM + DBT compared to FFDM alone or did not report on significance. There are considerable differences in the magnitude of this effect across the included studies (PPV_3 for FFDM + DBT ranges from 29.5% to 50%; for FFDM alone it ranges from 16.7% to 38.5%). Destounis et al. (2014) report a very large difference (50% compared to 16.7%); however statistical significance was not achieved. If Destounis et al.'s data is removed, the mean difference in PPV_3 between FFDM + DBT and FFDM alone is approximately 4% (although this is based on non-significant results). This indicates that on average for every 100 biopsies performed because of screening results, an additional four cases may be diagnosed with breast cancer using FFDM + DBT compared with FFDM alone.

There is some evidence to suggest that the accuracy of FFDM + DBT screening in correctly determining when biopsies should be performed may decrease with the age of women screened. Rafferty et al. (2017) is the only study to report the effect of age on PPV_2 and PPV_3. Their results show that FFDM + DBT accurately detected fewer cancers per number of biopsies performed than FFDM alone in women aged 70+ years. This is the opposite of results for women aged 40-69 years, which indicate that FFDM + DBT increases PPV_3 compared with FFDM alone, with statistically significant increases of 3.9% and 4.4% for women aged 40-49 and 50-59 years, respectively.

Two prospective trials (OTS and Malmö) reported PPV for biopsy rates but it was not clear from the articles whether the authors were reporting PPV for the number of cancers detected per the biopsies recommended (PPV_2) or performed (PPV_3). Regardless, the PPV for biopsy was similar for FFDM + DBT and FFDM alone, with neither comparison of PPV values achieving (or reporting) statistical significance.

Three studies reported data on PPV_3 for DBT + s2DM compared to either FFDM + DBT or FFDM alone, with considerable variation displayed in findings. Both Aujero et al. (2017) and Zuckerman et al. (2016) reported strong increases in PPV_3 for DBT + s2DM compared to either FFDM + DBT or FFDM alone. For example, Aujero et al. reported PPV_3 of 40.8% for DBT + s2DM compared to 28.5% and 22.3% for FFDM + DBT and FFDM alone respectively. Only a small difference is detected in Freer et al.'s 2017 study (a difference of 0.4% is reported).

Overall results on PPV_2 and PPV_3 indicate that FFDM + DBT was more accurate than FFDM alone when used as a basis for recommending or performing biopsies. PPV results for DBT + s2DM are also promising but present more varied effect size than results for FFDM + DBT; however, the direction of effect indicates that DBT + s2DM might be useful in terms of reducing false positives leading to either recall or biopsies.

Table 16: Studies reporting PPV rates for FFDM + DBT and FFDM alone

Study	Sample	Study type	PPV
FFDM + DBT compared to FFDM alone			
Prospective trials embedded in European population-based screening programs with biennial screening			
Skaane et al., 2013 (OTS trial)	12,631 women aged 50-69 years participating in the biennial Oslo breast screening program, with nine months follow-up.	Prospective, fully paired trial using Hologic Dimensions system with double reading	PPV: (*p*=.72) 28.5% FFDM alone 29.1% FFDM + DBT
Retrospective, American studies set in community-based radiology practices with annual screening			
Powell et al., 2017	FFDM + DBT: 2304 FFDM: 10,477	Retrospective observational data review of images generated with Hologic's Selenia + Dimensions systems	PPV_1: (*p*=.032) 5.6% FFDM + DBT 3% FFDM PPV_2: (*p*=.689) 29.5% FFDM + DBT 25.1% FFDM PPV_3: (*p*=.516) 29.5% FFDM + DBT 25.3% FFDM
Rafferty et al., 2017 (PROSPR consortium centres)	FFDM + DBT: 173,414 FFDM: 278,906	Retrospective, multicentre analysis of images taken using Hologic's Selenia Dimensions system	PPV_1: 40-49y: (*p*=.001) 3.4% FFDM + DBT 2.3% FFDM alone 50-59y: (*p*=.001) 6.0% FFDM + DBT 3.8% FFDM alone 60-69y: (*p*=.001) 10.3% FFDM + DBT 6.9% FFDM alone 70+: (*p*=.003) 12.6% FFDM + DBT 9.6% FFDM alone PV_3: 40-49y: (*p*=.006) 17.6% FFDM + DBT 13.7% FFDM alone 50-59y: (*p*=.012) 26.3% FFDM + DBT

Study	Sample	Study type	PPV
			21.9% FFDM alone 60-69y: (*p*=.077) 39.2% FFDM + DBT 35% FFDM alone 70+: (*p*=.55) 43.0% FFDM + DBT 45.1% FFDM alone
Conant et al., 2016	FFDM + DBT: 55,998 FFDM: 142,883	Retrospective analysis of data from three PROSPR consortium sites (NB mammography system used not stated)	PPV_1: (*p*=.0001) 6.4% FFDM + DBT 4.1% FFDM
McDonald et al., 2016	FFDM + DBT: 33,740 FFDM: 10,728 *12079 had one screen* *6293 had two screens* *3023 had three screens*	Retrospective review of mammography metrics from a single site over four years of screening with single reading using Hologic Dimensions system	PPV_1 (Year 0/Year3) (*p*=.02): 6.7% FFDM + DBT 4.4% FFDM
Lourenco et al., 2015	FFDM + DBT: 12,921 FFDM: 12,577	Retrospective review of two cohorts (DBT alone=2012/13, FFDM=2011/12), single reading with CAD. FFDM performed using GE Senographe series. DBT performed with Hologic Selenia Dimensions system.	PPV_3: (*p*=.21) 30.2% FFDM 23.8 DBT
McDonald et al., 2015	FFDM + DBT (Prevalent): 1859 FFDM + DBT (Incident): 9524 FFDM (Prevalent): 1204 FFDM (Incident): 13,712	Observational study using Hologic Dimensions system for women presenting for mammographic screening at a single institution *NB Data comes from McCarthy et al., 2015*	PPV_1: prevalent (*p*=.25) 3.7% FFDM + DBT 2.0% FFDM PPV_1: incident (*p*=.09) 6.9% FFDM + DBT 5.1% FFDM PPV_2: prevalent (*p*=.81) 12.8% FFDM + DBT 14.5% FFDM PPV_2: incident (*p*=.48) 27.6% FFDM + DBT 24.6% FFDM PPV_3: prevalent (*p*=.84) 14.1% FFDM + DBT 15.6% FFDM PPV_3: incident (*p*=.65) 28.7% FFDM + DBT 26.6% FFDM (incident)
Destounis et al., 2014	524 women aged >30 years (mean age 59 years) including women with a history of BC	Retrospective review of images with double reading. FFDM system was Hologic Selenia or Dimensions, GE Senographe Essential or Fuji CRm. DBT system was Hologic Selenia Dimensions.	PPV_3: 16.7% FFDM 50.0% FFDM + DBT
Friedewald et al., 2014 (PROSPR consortium centres)	FFDM + DBT: 173,663 FFDM: 281,187	Retrospective review with single reader using data from 13 centres all using Hologic Selenia Dimensions systems	Mean PPV_1: (*p*<.001) 6.1% FFDM + DBT 4.1% FFDM Mean PPV_3: (*p*<.001) 29.2% FFDM + DBT 24.2% FFDM
Greenburg et al., 2014	FFDM + DBT: 20,943 FFDM: 38,674 No differences in study arms by age, ethnicity, family history of BC, or prevalence or incidence screening	Retrospective review of mammography outcomes at a multi-site radiology practice using Hologic Selenia or Selenia Dimensions systems	PPV_1: (*p*=.0003) 4.6% FFDM + DBT 3.0% FFDM PPV_3: 22.7% FFDM + DBT 21.5% FFDM
McCarthy et al., 2014	FFDM + DBT: 15,571	Observational study using Hologic	PPV_1: (*p*=.047)

Study	Sample	Study type	PPV
	FFDDM: 10,728	Dimensions system for women presenting for mammographic screening at a single institution	6.2% FFDM + DBT 4.4% FFDM PPV_2: 24.7% FFDM + DBT 22.4% FFDM PPV_3: 25.4% FFDM + DBT 24.7% FFDM
Rose et al., 2013	FFDM + DBT: 9,499 FFDDM: 23,355	Retrospective observational study at a multi-site community-based breast screening centre before and after implementation of DBT using Hologic's Selenia and Dimensions systems	PPV_1: 10.1% FFDM + DBT 4.7% FFDM Average PPV_3: 39.8% FFDM + DBT 26.5% FFDM
Retrospective studies from other settings			
Pan et al., 2017	No specific description of the sample provided	Retrospective analysis comparing screening outcomes before/after implementation of DBT (Hologic Dimensions system) from a single hospital site to national data from Taiwan's National Cancer Registry	Average PPV_1: 10.1% FFDM + DBT 6.1% FFDM Average PPV_2: 33.27% FFDM + DBT 31.0% FFDM Average PPV_3: 38.47% FFDM + DBT 38.5% FFDM

Table 17: Studies reporting PPV rates for DBT + s2DM, FFDM + DBT and FFDM alone

Study	Sample	Study type	PPV
Prospective trials embedded in European population-based screening programs with biennial screening			
Skaane et al., 2014 (OTS trial)	12,270 screens from 24,901 Norwegian women aged 50-69 years (mean age 59.2 years)	Prospective, fully paired trial using Selenia Dimensions system with C-view for FFDM + DBT with double reading and two study periods (one using C-view, one using an earlier software. Only rates using C-view are reported here)	Study period 1: PPV_1 (p=.61) 30.3% DBT + s2DM 28.5% FFDM + DBT Study period 2: PPV_1 (p=.47) 34.9% DBT + s2DM 32.1% FFDM + DBT
Retrospective, American studies set in community-based radiology practices with annual screening			
Aujero et al., 2017	Mammograms from a single USA practice: 16,173 mammograms with DBT + s2DM; 30,561 mammograms with FFDM + DBT; 32,076 mammograms with FFDM alone	Retrospective observational study with single reading using Selenia Dimensions system with C-view	PPV_1: (p=.001) 14.3% DBT + s2DM 10.9% FFDM + DBT 6% FFDM alone PPV_2: (p=.01) 39.3% DBT + s2DM 26.3% FFDM + DBT 20.9% FFDM alone PPV_3: (p=.001) 40.8% DBT + s2DM 28.5% FFDM + DBT 22.2% FFDM alone
Freer et al., 2017	31,979 women receiving a screening mammogram a single USA practice between 10/2013–12/2015 (9525 women screened with DBT + s2DM; 1019 screened with FFDM + DBT; 21,435 screened with FFDM alone	Retrospective analysis using Hologic Selenia and Dimensions systems with C-view	*Adjusted[3] PPV* PPV_1: 9.1% DBT + s2DM 8.1% FFDM + DBT 6.2% FFDM alone PPV_2: (p=.054) 36.4% FFDM + DBT 40.3% DBT + s2DM

Study	Sample	Study type	PPV
			30.9% FFDM alone PPV_3: (*p*=.53) 36.7% FFDM + DBT 36.3% DBT + s2DM 31% FFDM alone
Zuckerman et al., 2016	15,571 American women screened with FFDM + DBT and 5366 women screened with DBT + s2DM	Observational study set in a community screening setting using Hologic Dimensions system	PPV_1: (*p*=.58) 7.1% DBT + s2DM 6.2% FFDM + DBT PPV_2: (*p*=.054) 35.5% DBT + s2DM 24.7% FFDM + DBT PPV_3: (*p*=.53) 38.6% DBT + s2DM 27% FFDM + DBT

Table 18: Studies reporting PPV rates for DBT_{MLO} plus/compared to DM_{CC} or FFDM alone

Study	Sample	Study type	PPV
Prospective trials embedded in European population-based screening programs with biennial screening			
Lång et al., 2016A (Malmö)	7,500 women aged 40-74 years participating in the Swedish screening program	Prospective, fully-paired one-arm, single-institution study of DBT_{MLO} alone versus FFDM and DBT_{MLO} and DM_{CC} using Siemens Mammomat	PPV: 24% for FFDM and DBT No p-value provided

Summary

The literature is not settled about the association between DBT and recall rates. It shows that overall recall rates can be significantly reduced when using FFDM + DBT and DBT + s2DM compared with FFDM alone, but there is strong variance in the data reported. Differences in reading and arbitration protocols used to determine which women to recall from screening may account for some of the inconsistency between results reported by the large prospective trials. Double reading (either by two radiologists or through an arbitration process) increased recall rate in some prospective trials (but reduced the false positive rate) or reduced both recall rate and false positives in others. Information from the smaller retrospective studies (most of which used single reading strategies reported that recall rate was reduced with the addition of DBT to FFDM). The OTS trial reported two rates: a lower recall rate with pre-arbitration and a higher recall rate with post-arbitration for FFDM + DBT compared to FFDM alone. Results from STORM (double reading with recall if either radiologist reported a positive finding) noted that recall rates and false positives were lower for FFDM + DBT. Preliminary results on the Malmö trial (Lång et al., 2016A) have reported a statistically significant increase in overall recall rates when using DBT_{MLO} compared with FFDM. Factors that may affect the development of the evidence around the association include readers gaining more experience and knowledge around the new technology, and as prior DBT images become available for comparison at incident screening with DBT (thus providing more information to readers).

There is less evidence currently available on the effect of DBT + s2DM on the rate of false positive recalls. One retrospective American study (Aujero et al., 2017) showed that false positive recall rates are significantly reduced when using DBT + s2DM compared with both FFDM + DBT and FFDM alone. However, results from a large prospective trial (STORM, Bernardi et al.,2016) showed a false positive recall rate for DBT + s2DM that was greater than for FFDM + DBT and FFDM alone. It is possible that this resulted from an early incorporation of s2DM into

real world screening practice for the first time without readers having previous experience with s2DM images relative to their expertise with FFDM imaging in the screening program involved in the trial. Secondary analysis from the STORM-2 trial indicated that false positive recall rates for FFDM + DBT and DBT + s2DM are significantly lower than those for FFDM. The false positive recall rate of DBT + s2DM was still slightly higher than that for FFDM + DBT. These results may be indicative of an increased knowledge in the use of FFDM + DBT, with some interpretation issues still present for s2DM. Interim results from the Malmö trial reported a reduction in the false positive recall rate of screens using DBT over the first 1.5 years, which indicates that false positive recall will be associated with a learning curve in interpretation.

3.3. The impact of DBT for different population groups: breast density and age

The literature reports evidence across a range of clinical outcomes and performance metrics that DBT (used in a range of different screening strategies) may perform differentially for different population groups. Much of this evidence focuses on either performance of DBT on different age groups or by breast density. More information on the role of other adjunct screening modalities for women with higher breast density is discussed in *Allen + Clarke*'s literature review on breast density and screening (also commissioned by the Department of Health).

3.3.1. Why are these two population groups important?

Breasts are made up of fat and fibroglandular (non-fatty) tissue with the composition of breast tissue varying between women. Breast density assessments are made by looking for mammographic parenchymal patterns.

Bilateral FFDM is the current "gold standard" screening test for early detection of breast cancer in most national breast cancer screening programs including the BSA program (which are designed for average-risk, asymptomatic women). On a mammogram, fatty tissue appears black while the remaining breast tissue appears white or radiographically 'dense', with the relative amount of fibroglandular tissue areas on a mammogram referred to as breast (or mammographic) density (i.e., heterogeneously or extremely dense breasts with little fat). In FFDM, sensitivity for cancer detection can be lower for women with more dense breasts because, due to their similar X-ray attenuation properties, cancers may also appear as white areas on mammograms (*NB* fat has a lower X-ray attenuation and appears darker, making cancers in less dense breasts easier to see). Also, certain dense breast structures can be superimposed which can mask cancers or which can make areas of normal tissue appear suspicious for cancer. Together, conspicuity is reduced making it difficult for readers to clearly differentiate between normal tissue and malignancy. This makes some cancers more difficult to detect in some women with more dense breasts and can interfere with the interpretation of mammograms.

Breast density declines with age, with international research indicating more than half of women under the age of 50 years have more dense breasts; for women over 50 years of age about one third have more dense breasts (Berg et al., 2008 cited in Coop et al., 2016). Women can start participating in the BSA program at 45 years (by invitation from 50 years). As women aged under 50 years are more likely to have more dense breasts and CDR is generally higher in women having prevalent (i.e., first) mammograms compared to later incident screening, breast

density and age are two key population sub-group parameters that should be explored when considering DBT's potential impact as a screening strategy.

We identified seven articles (generated from 11 studies) that reported on CDR, overall recall and false positive rates stratified by age and/or breast density, including one literature review (Houssami and Turner, 2016). No articles reported on differences by ethnicity or provided other stratification. Maxwell et al. (2016), who completed the only RCT informing this literature review, did not complete stratification by age or breast density. Key results are reported in Tables 19 – 24.

3.3.2. DBT and women with more dense breasts

CDR

Coop et al. (2016) cited evidence that FFDM + DBT reduces the effects of overlap and can lead to increases in overall CDR and invasive CDR for women with heterogeneously or extremely dense breasts. They noted that the greatest change in CDR was experienced by younger women with heterogeneously or extremely dense breasts (BIRADS 3-4), although no further discussion of the evidence underpinning this statement is provided. Hodgson et al. did not analyse CDR results by population subgroup (either age or breast density).

In July 2016, Houssami & Turner (2016) completed a rapid evidence review investigating incremental CDR of FFDM + DBT in the screening of women with more dense breasts. While not a systematic review, this rapid review provided pooled analysis of 10,188 women across eight studies (see Table 19, overleaf). As with other reviews, the authors considered prospective and retrospective studies separately. Houssami & Turner reported differences in the incremental CDR between the results reported from larger prospective studies compared to retrospective studies. Most of the results (including from the STORM trial and all the retrospective studies) reported on FFDM + DBT compared to FFDM alone. Using this screening strategy, prospective trials reported increases in the number of extra cancers identified using FFDM + DBT compared to those detected with FFDM alone (i.e., incremental CDR). They reported an increase in incremental CDR of between 2.5 and 4.0 cancers per 1000 screening examinations (pooled analysis = 3.9 cancers per 1000 screening examinations) for FFDM + DBT compared to FFDM.

Study **Design** **Quality**	**Results: Incremental CDR attributed to FFDM + DBT compared to FFDM alone**
Houssami & Turner, 2016 Rapid review	Pooled analysis from: *Prospective studies:* (10,188 women): 3.9 (95%CI: 2.7-5.1, *p*<.001) *Retrospective studies:* (281,044 screening events): 1.4 (95%CI: 0.9-2.0, *p*<.001)

While using a sequential approach to screening[15], Lång et al. (2016) reported that DBT alone detected more cancers in both more dense and more fatty breasts compared to FFDM alone. Lång et al. reported that this may mean that increases in CDR are not only due to improved conspicuity seen with DBT compared to FFDM alone. Other prospective trials (STORM and OTS) also reported increased CDR for all women regardless of BIRADS classification (eg, Skaane et al.,

[15] DBT_{MLO} followed by DBT_{CC} followed by prior mammogram images where available compared to FFDM followed by prior mammogram images where available and breast density.

2013A, reported comparable CDR for BIRADS 1-2 compared to BIRADS 3-4; Ciatto et al., 2013, reported CDR of 2.8 cancers per 1000 screening examinations for women with BIRADS 1-2 compared to 2.5 per 1000 screening examinations for BIRADS 3-4).

Results for the retrospective studies varied between incremental detection of 1.4 and 2.1 cancers per 1000 screening examinations (pooled analysis = 1.4 cancers per 1000 screening examinations), favouring FFDM + DBT compared to FFDM alone. While these rates reported the same direction of association, they were generally (but not always) higher than the overall CDR reported for each study.

This review's findings for CDR stratified by breast density may present results that are surprising, given that DBT improves conspicuity and should, in theory, provide improved images for women with more dense breasts which could lead to increased CDR for this population. We note that the use of BIRADS to assess breast density can result in unreliable allocation to BIRADS category 2 and 3. This is because density classification can be affected by factors like hormone levels, genetic factors, parity, use of oestrogen, place in menstrual cycle, use of tamoxifen, weight and inter/intra reader variability. It is possible for women to be classified as having non-dense breasts in one mammogram but be reclassified to having more dense breasts in the next mammogram (and vice versa). This creates a level of unreliability that could account for the smaller-than-expected incremental increase in CDR between women with more dense or less dense breasts. It may be that density classifications which report CDR, recall and false positives by 25th percentile (very dense) and 75th percentile (very fatty) could result in clearer (and possibly truer) incremental differences in CDR by density; however, the research undertaken to date does not make this comparison (i.e., between BIRADS 1 and BIRADS 4).

Table 19: Studies reporting on breast density (from Houssami & Turner, 2016)

Study	Study participants with dense breasts (BIRADS 3 or 4)	Study type	Increase in cancers detected *All rates from Houssami & Turner, 2016*
Prospective trials embedded in European population-based screening programs with biennial screening			
Bernardi et al., 2016 (STORM-2)	2592 women Total sample: 9677 women	Prospective, fully paired trial comparing FFDM, FFDM + DBT and DBT + s2DM using double reading using Hologic's Selenia Dimensions system *NB Incremental increase reported for FFDM + DBT vs FFDM alone*	Increase of 5.4
Lång et al., 2016A (Malmö)	3150 women Total sample: 7500 women	Prospective, fully paired trial comparing a sequential approach to screening DBT_{MLO} followed by DM_{CC} compared to FFDM using double reading	Increase of 3.8
Tagliafico et al., 2016 (ASTOUND trial)	3231 Italian women, median age = 51y (44 – 78 years)	Prospective comparative trial comparing ultrasound and DBT performance for women with dense breasts, using Hologic Dimensions system using single reading	Increase of 4.0 (1.8 to 6.2)
Ciatto et al., 2013 (STORM)	1215 women Total sample: 7294 women	Prospective, fully paired trial comparing FFDM alone to FFDM + DBT using double reading using Hologic's Selenia Dimensions system	Increase of 2.5
Retrospective, American studies set in community-based radiology practices with annual screening			
Rafferty et al., 2017	FFDM + DBT: 84243 FFDM: 131,996	Review of screening clinical outcomes and performance metrics comparing FFDM alone to FFDM + DBT	Increase of 1.4
Conant et al., 2016	FFDM + DBT: 9265 FFDM: 35320	Retrospective analysis of data from three PROSPR consortium sites comparing FFDM alone to FFDM + DBT (NB mammography system used not stated)	Increase of 2.1

McCarthy et al., 2014	FFDM + DBT: 5056 FFDDM: 3489	Observational study using Hologic Dimensions system for women presenting for mammographic screening at a single institution comparing FFDM alone to FFDM + DBT	Increase of 1.8
Rose et al., 2013	FFDM + DBT: 4666 FFDM: 7009	Observational study comparing FFDM alone to FFDM + DBT following practice implementation of DBT	Increase of 1.4

Recall rates and false positives

Ten studies reported in Houssami and Turner (2016) discussed overall recall rates and false positive recall rates and stratified results by breast density.

Overall recall rates

Study Design Quality	Results: recall rates for different breast densities
Houssami & Turner, 2016 Rapid review	Pooled analysis from: *Prospective studies:* (not performed) *Retrospective studies:* 2.33% reduction in recall rates for women with dense breasts (p<.001)

Houssami & Turner (2016) provided pooled analysis (separated into prospective and retrospective studies) on recall rates for women with dense breasts (BIRADS 3 or 4). Houssami & Turner's meta-analysis of recall rates data from the retrospective studies (115,098 examinations using DBT compared to 186,797 FFDM screening examinations) showed a statistically significant difference of 23.3 fewer recalls per 1000 screening examinations for FFDM + DBT compared to FFDM for women with dense breasts. FFDM + DBT resulted in a significant reduction in recall rates for women with more dense breasts screened with DBT in comparison with FFDM alone. Houssami and Turner did not complete any further analysis looking at differences between women with more dense breasts screened by FFDM + DBT or FFDM compared to women with less dense breasts (BIRADS 1-2). Data on individual studies is reporting in Table 20 (below).

Table 20: Studies reporting on overall recall rates by breast density (From Houssami & Turner, 2016)

Study	**Study participants with dense breasts (BIRADS 3 or 4)**	**Study type**	**FFDM + DBT** (95%CI; *p*-value)	**FFDM alone** (95%CI; *p*-value)	**Difference between recall rates for dense breasts / 1000 screens** (95%CI; *p*-value) *All rates from Houssami & Turner, 2016*
Retrospective, American studies set in community-based radiology practices with annual screening					
Conant et al., 2016	FFDM + DBT: 9265 FFDM: 35320	Retrospective analysis of data from three PROSPR consortium sites	Overall: 8.7% Non-dense breasts: 7.4% Dense breasts: 10.3% (p<.0001)	Overall: 10.4% Non-dense breasts: 9.1% Dense breasts: 12.6% (p<.0001)	22.1 decrease (p<.0001)
Rafferty et al., 2016	FFDM + DBT: 84243 FFDM: 131,996	Review of screening performance metrics	92/1000 screens	106/1000 screens	18.4 decrease

McCarthy et al., 2014	FFDM + DBT: 5056 FFDDM: 3489	Observational study using Hologic Dimensions system for women	Non-dense breasts: 7.8% Dense breasts: 10.8%	Non-dense breasts: 9.2% (p<.001) Dense breasts: 12.8% (p<.006)	19.4 decrease
Rose et al., 2013	FFDM + DBT: 4666 FFDM: 7009	Observational study following practice implementation of DBT	Overall: 5.5% BIRADS 1: 2.7% BIRADS 2: 4.3% BIRADS 3: 6.6% BIRADS 4: 9.0%	Overall: 8.7% BIRADS 1: 4.6% BIRADS 2: 7.2% BIRADS 3: 10.2% BIRADS 4: 13.3%	36.8 decrease for women with dense breasts

Lång et al. (2016B) reported interim results from the Malmö trial (using Siemens' Mammomat Inspirations system) and noted that women recalled following FFDM + DBT tended to be slightly younger (the average age was 51 years) and were more likely to have more dense breasts compared to women recalled on FFDM or DBT_{MLO} alone (who were slightly older at an average age of 55 years). Women recalled in the DBT_{MLO} reading arm tended to have slightly less dense breasts compared with women recalled in the DM_{CC} arm of the study. This report did not provide any specific values for recall rates stratified by age and breast density, beyond noting these observations.

Allen + Clarke's review identified three retrospective studies (Sharpe et al., 2016; McDonald et al., 2015; Haas et al., 2013) that stratified recall rates by breast density, but which were not discussed by Houssami & Turner. Detailed results are included in Table 21 (below). While all these studies show statistically significant reductions in recall rate for all women when screened with FFDM + DBT compared to FFDM alone, the general direction of effect is that the use of FFDM + DBT resulted in even lower recall rates for women with dense breasts when compared with FFDM. Sharpe et al. (2016) reported that overall recall rates reduced most of all for women with dense breasts (BIRADS 3 to 4) when screened with FFDM + DBT compared with FFDM alone. For women with extremely dense breasts, overall recall rates decreased by 27.45% (p=.0429) when screened with FFDM + DBT. For women with predominantly fatty breasts, overall recall rates decreased by 18.68% (p<.0001).

McDonald et al. (2015) reported a decrease in overall recall rates between FFDM + DBT compared with FFDM of 13.1% (p=.01) for women with non-dense breasts, and 15.4% for women with dense breasts (p=.01). Haas et al. (2013) also reported a reduction in recall rates for all BIRADS categories for FFDM + DBT compared to FFDM alone. The greatest reductions in recall rates occurred for women with dense breasts screened with FFDM + DBT rather than FFDM alone. Women with the densest breasts (BIRADS 4) had a reduction in overall recall rates of 57.3% (p<.01) and women with the fattiest breasts (BIRADS 1) had a reduction in recall rate of 30% (p=.12).

Table 21: Retrospective studies reporting on overall recall rates by breast density

Study	**Study participants with dense breasts (BIRADS 3 or 4)**	**FFDM + DBT** (95%CI; p-value)	**FFDM alone** (95%CI; p-value)	**Difference between recall rates for different age groups** (95%CI; p-value)
Sharpe et al., 2016	FFDM + DBT: 5703 FFDM: 80,149	Overall: 6.10% BIRADS 1:4.87% BIRADS 2: 6.39% BIRADS 3: 7.33% BIRADS 4: 4.74%	Overall: 7.51% BIRADS 1: 5.5% BIRADS 2: 7.03% BIRADS 3:9.31% BIRADS 4: 6.54%	18.68% decrease (p<.0001) 11.38% decrease (p=.3914) 9.04% decrease (p=.3335) 21.25% decrease (p=.0048) 27.45% decrease (p=.0429)

McDonald et al., 2015	FFDM + DBT: 33,740 FFDM: 10,728 *12079 had one screen* *6293 had two screens* *3023 had three screens*	Overall: 16% BIRADS 1 & 2: 6.8% BIRADS 3 & 4: 9.9%	Overall: 20.5% BIRADS 1 & 2: 7.8% BIRADS 3 & 4: 11.7%	22% decrease (p=.002) 13.1% decrease (p=.01) 15.4% decrease (p=.01)
Haas et al., 2013	FFDM + DBT: 6100 FFDM alone: 7058	BIRADS 1: 5% BIRADS 2: 7.9% BIRADS 3: 10.2% BIRADS 4: 6.7%	BIRADS 1: 7.2% BIRADS 2: 10.6% BIRADS 3: 16.7% BIRADS 4: 15.6%	30% decrease (p=.12) 25% decrease (p<.01) 39.4% decrease (p<.01) 57.3% decrease (p<.01)

False positive recall rates

Two of the prospective studies that Houssami & Turner (2016) cited in their rapid review reported on false positive recall rates from the two STORM trials (see Table 22, below). Results are inconsistent.

- For STORM, Ciatto et al. (2013) reported a decrease in false positive recall rates of 26 per 1000 screening examinations for women with dense breasts (BIRADS 3-4) when screened with FFDM + DBT compared to FFDM alone. This compares to a whole population decrease in false recall of 9.3 false recalls per 1000 screening examinations.
- For STORM-2, Bernardi et al. (2016) reported an increase in false positive recall rates when screened with FFDM + DBT (5.01% for women with dense breasts - BIRADS 3-4 - compared to 3.95% when screened with FFDM alone). Bernardi et al. reported a false positive recall rate of 3.23% for FFDM and a false positive rate of 3.6% for FFDM + DBT for women with less dense breasts (BIRADS 1-2). For women with more dense breasts (BIRADS 3-4), the authors reported a false positive rate of 3.95% for FFDM, and 5.01% for FFDM + DBT. The authors of the review did not discuss any potential reasons for this disparity.

Also, from STORM-2, Bernardi et al. (2016) reported stratified false positive recall rates for breast density for DBT + s2DM. For more dense breasts, DBT + s2DM resulted in a 2.11% increase in false positive recall rates when compared to FFDM (p<.0001). FFDM + DBT resulted in a 1.06% increase when compared to FFDM (p=.0016). For women with less dense breasts, DBT + s2DM and FFDM + DBT still resulted in higher rates of false positive recall than FFDM alone, but to a lesser extent. The authors concluded that integrated FFDM + DBT and DBT + s2DM had higher rates of false positive recall than FFDM alone for women with more dense breasts. However, they qualified that by noting that the false positive recall rates might be decreased through further experience and repeated screening with DBT.

Table 22: STORM-2 false positive recall results from Bernardi et al. (2016) stratified by age and breast density

Stratification	**DBT + s2DM** (95%CI; *p*-value)	**FFDM + DBT** (95%CI; *p*-value)	**FFDM alone** (95%CI; *p*-value)	**Difference compared with FFDM alone** (95%CI; *p*-value)
Breast density				
Dense breasts	6.07%	5.01% (4.20-5.93; p=.0016)	3.95% (3.23-4.78)	DBT + s2DM: 2.11% increase (p<.0001) FFDM + DBT: 1.06% increase (p=.0016)
Non-dense breasts	3.87%	3.6% (3.14-40.6; p=.051)	3.23% (2.83-3.67)	DBT + s2DM: 0.64% increase (p=.051) FFDM + DBT: 0.37% increase (p=.014)
Age				
Less than 60 years	5.01%	4.31%	3.94%	DBT + s2DM: 1.07% increase (p=.00085) FFDM + DBT: 0.37% increase (p=.10)
60 years or older	3.63%	3.48%	2.65%	DBT + s2DM: 0.98% increase (p=.0035) FFDM + DBT: 0.82% increase (p=.00054)

Lång et al. (2016B) reported different false positive recall rates according to breast density (based on BIRADS classification) of the women recalled. For DBT_{MLO}, 40.5% of false positive

recalls were in women with less dense breasts, and 59.5% of false positive recalls were in women with more dense breasts. For the FFDM alone reading arm, 27.7% of false positive recalls were in for women less dense breasts, and 63% of false positive recalls were for women with more dense breasts. For images read with DM_{CC} + DBT_{MLO}, 33.3% of falsely recalled women had less dense breasts, and 67% had more dense breasts. These results are consistent with Bernardi et al.'s (2016) findings that FFDM + DBT resulted in higher false positive recalls in general and especially for women with dense breasts.

In another study using the interim results from the Malmö trial, Rosso et al., (2015) reported false positive recall rates for women with more dense breasts, finding that the overall false positive rate was lower with the use of FFDM compared to DBT, and the false positive rates for both methods increase with breast density in the same manner as reported by Lång et al. (2016B).

3.3.3. Younger and older women

Eight articles (generated from eight studies) reported CDR and recall rates stratified by age.

Age stratification completed in studies included in this literature review was not done with consistent age banding. Some studies reported results separated into 10-year age bands (Rafferty et al., 2017; Conant et al., 2016). Other studies reported results in groups of women aged under or over 60 years or under or over 50 years (Bernardi et al., 2016; McCarthy et al., 2016; Ciatto et al., 2013). This makes comparison of results challenging and limits our ability to consider the association between age, breast density and screening strategy (whether DBT is used alone or as an adjunct screen to FFDM).

Key results are reported in Table 23 and Table 24 (see pages 79 and 80).

CDR

Overall, the literature on FFDM + DBT compared to FFDM alone for all women reported a greater increase in CDR. The literature also reports on variation by age.

In the STORM trial, Ciatto et al. (2013) reported an elevated incremental CDR for all women but the incremental CDR was much higher for women aged over 60 years compared to younger women (4.0 cancers per 1000 screening examinations compared to 1.7; $p=.016$). This association of elevated overall CDR and increased CDR for older women is seen in other studies as well. For example, in their larger retrospective, multicentre analysis, Rafferty et al. (2017) reported statistically significant incremental CDR per 1000 screening examinations of 0.9, 1.4 and 1.7 for women aged 40-49 years, 50-59 years and 60-69 years respectively. Conant et al. (2016) reported the same trend although their results did not achieve significance.

Conversely, one study (McCarthy et al., 2014) reported age stratification for CDR results and found increased CDR for younger women (<50 years) compared to older women (>50 years), although the increased CDR for younger women is not much higher than the result reported for women aged over 50 years. The authors note that this result may be affected by a small sample size (27 cancers detected in 4365 women).

In their retrospective study, Haas et al. (2013) did not report directly on age or breast density but investigated differences in CDR for women with an increased or baseline risk of breast cancer based on personal history of breast cancer or a first-degree relative with breast cancer. The authors reported a non-significant increase in CDR for women with an elevated risk of

breast cancer with the use of FFDM + DBT compared to FFDM alone (8.6 per 1000 screening examinations compared to 7.9 for women with a baseline risk profile, *p*=.83).

Using data from the STORM-2 trial, Bernardi et al. described age differences for women aged <60 years and >60 years by three screening strategies (DBT + s2DM, FFDM + DBT, and FFDM alone). Results were not reported separately for women aged under 50 years. The authors found that CDR was higher for women aged over 60 years with all screening strategies; however, there is considerable overlap between the 95%CI reported for the comparisons between DBT + s2DM for each age group. The authors also noted a statistically significant increase in incremental CDR for DBT + s2DM compared to FFDM alone in younger women. This equated to an additional 3.3 cancers per 1000 screening examinations (95%CI 1.9-5.2; *p*=.0001). Statistical significance was not achieved for women aged over 60 years: incremental CDR was 1.3 cancers per 1000 screening examinations (95%CI -0.6-3.3; *p*=.23). This finding suggests that DBT + s2DM increases CDR for all women but that DBT + S2DM may have particularly strong impact on CDR for younger women. Given the difference of this result compared to other studies, further research is needed.

Study Design Quality	**Results: CDR stratified by age and BIRADS for density**
Bernardi et al., 2016 Trial (STORM-2) High quality	Women aged under 60y DBT + s2DM: 7.0 (5.0-9.5) FFDM + DBT: 6.3 (4.4-8.7) FFDM alone: 3.7 (2.3-5.6) Women aged over 60y DBT + s2DM: 10.2 (7.3-13.8) FFDM + DBT: 11.7 (8.6-15.6) FFDM alone: 10.2 (7.3-13.8) Women with BIRADS 1 or 2 DBT + s2DM: 6.9 (5.1-9.1) FFDM + DBT: 6.8 (5.0-9.0) FFDM alone: 5.8 (4.2-7.8) Women with BIRADS 3 or 4 DBT + s2DM: 13.9 (9.7-19.2) FFDM + DBT: 13.1 (9.1-18.3) FFDM alone: 7.7 (4.7-11.9)

No data on CDR for different population subgroups is reported in Lång et al.'s (2016A) interim analysis; however, all CDR is expected to be higher at prevalent screening and lower at later incident screening. Lång et al. noted that the CDR finding may be higher than expected in a general screening population because of the high percentage of study participants undergoing a prevalent screen (20% of the sample, who would either be younger women or new arrivals to Sweden). It is important to follow this in the Malmö trial's final analysis and to consider what, if any, impact the different screening strategies have on whether breast cancer detection has a greater increase in younger or older women.

Table 23: Studies reporting on CDR results stratified by age and/or breast density

Study	Sample	Study type	Stratification by age			Stratification by breast density[1]		
			FFDM + DBT All CDR per 1000 screening examinations (95%CI; p-value)	FFDM alone All CDR per 1000 screening examinations (95%CI; p-value)	Incremental CDR per 1000 screening examinations (95%CI; p-value)	FFDM + DBT All CDR per 1000 screening examinations (95%CI; p-value)	FFDM alone All CDR per 1000 screening examinations (95%CI; p-value)	Incremental CDR per 1000 screening examinations (95%CI; p-value)
Prospective trials embedded in European population-based screening programs with biennial screening								
Ciatto et al., 2013 (STORM) *Main article*	<60y: 4044 >60y:3250 BIRADS 1-2: 6079 BIRADS 3-4: 1215	Prospective, fully paired trial using Hologic Selenia Dimensions systems in combination mode	<60y: 6.7 (4.4-9.7) >60: 9.8 (6.7-13.9)	<60: 4.9 (3-7.6) >60: 5.8 (3.5-9.1)	<60: 1.7 (0.7-3.6, *p*=.016) >60: 4.0 (2.1-6.8, *p*<.0001)	BIRADS 1-2: 8.4 (6.3-11.0) BIRADS 3-4: 6.6 (4.1-18.6)	BIRADS 1-2: 5.6 (3.5-7.8) BIRADS 3-4: 4.1 (3.1-9.6)	BIRADS 1-2: 2.8 (1.6-4.5, *p<.0001*) BIRADS 3-4: 2.5 (0.5-7.2, *p*=.25)
Retrospective, American studies set in community-based radiology practices with annual screening								
Rafferty et al., 2017	40-49y: 127,276 50-59y: 144,344 60-69y: 107,233 >70y: 62,576	Retrospective, multicentre analysis of images taken using Hologic's Selenia Dimensions system	40-49y: 3.8 (3.1-4.4 50-59y: 5.0 (4.5-5.6 60-69y: 7.4 (6.4-8.3 >70y: 8.2 (6.6-9.7)	40-49y: 2.9 (2.3-3.5) 50-59y: 3.6 (3.1-4.1) 60-69y: 5.7 (4.9-6.5) >70y: 7.0 (5.7-8.3)	0.9 (*p*=.011) 1.4 (*p*=.001) 1.7 (*p*=.001) 1.2 (*p*=.1)	NR	NR	NR
Conant et al., 2016	40-49y: 55,823 50-74y: 143,508 BIRADS 1-2: 117,596 BIRADS 3-4: 65,436 BIRADS unknown: 15819	Retrospective analysis of data from three PROSPR consortium sites (NB mammography system used not stated)	40-49y: 4.7 50-74y: 6.5	40-49y: 2.9 50-74y: 5.0	NR	BIRADS 1-2: 5.3 BIRADS 3-4: 6.8	BIRADS 1-2: 4.1 BIRADS 3-4: 4.7	NR
Starikov et al., 2016	<u>FFDM + DBT</u> BIRADS 1-2: 195 BIRADS 3-4: 1875 <u>FFDM alone</u> BIRADS 1-2: 5040 BIRADS 3-4: 7117	Retrospective observational case-control study (System not stated)	NR	NR	NR	BIRADS 1-2: 5.1 (*p*=.45) BIRADS 3-4: 5.3 (*p*=.35)	BIRADS 1-2: 2.4 BIRADS 3-4: 3.8	NR
McCarthy et al., 2014	<50y: 7910 >50y: 18,389 BIRADS 1-2: 17,754 BIRADS 3-4: 8545	Observational study using Hologic Dimensions system for women presenting for mammographic screening at a single institution	<50y: 5.7 (3.5-7.9, *p*=.022) >50y: 5.4 (4.0-6.7, *p*=.84)	<50y: 2.2 (0.6-3.8) >50y: 5.6 (3.9-7.2)	NR	BIRADS 1-2: 4.8 (3.4-6.1, *p*=73) BIRADS 3-4: 6.9 (4.6-9.2, *p*=.33)	BIRADS 1-2: 4.3 (2.8-5.8) BIRADS 3-4: 5.2 (2.8-7.5)	NR

NR = not reported

Recall rates and false positives

While the RCT completed by Maxwell et al. (2017) did not report on stratification by age, this paper is included in this section because the study population were aged under 50 years. Maxwell et al. reported no difference between the FFDM + DBT compared to FFDM alone for overall recall rates for women between the ages of 40 and 49 (the only women included in this study). The overall recall rate for FFDM alone was 2.8%, and the overall recall rate for FFDM + DBT was 2.7% (no confidence intervals or *p*-values were provided for overall recall rates). Maxwell et al. hypothesised that the system of consensus/arbitration, whereby women are not automatically recalled based on the most suspicious opinion, may have contributed to keeping the recall rate so low in both study arms. In addition, because the study sample was women aged 40-49 years, no comparative analysis with other age groups was completed within this study. Maxwell et al. (2017) reported no statistically significant difference between the two screening groups for false positive recall rates. The false positive recall rate for FFDM alone was 2.4% (95%CI 1.7-3.4), and for FFDM + DBT was 2.2% (95%CI 1.9-3.8); FFDM + DBT vs. FFDM *p*=.89).

Conant et al. (2016) stratified recall rates by age, finding that DBT decreased recall rates more significantly for younger women aged less than 50 years compared with FFDM alone. Sharpe et al. (2016) reported that overall recall rates reduced most significantly for women in their fifth or seventh decades. Haas et al. (2013) reported that the greatest reductions in recall rates occurred for women younger than 50 years. McCarthy et al. (2014) also reported extensively on recall rates stratified by age. Rafferty et al. (2017) also found the greatest reduction in recall rates in women aged 40-49 years. Overall, DBT had a statistically lower recall rate irrespective of age, however, the impact was greater in women aged 50 years and older. The results of these retrospective studies are summarised in Table 24 (below).

Table 24: Retrospective studies reporting on overall recall rates by age

Study	**Study participants with dense breasts (BIRADS 3 or 4)**	**Study type**	**FFDM + DBT** (95%CI; *p*-value)	**FFDM alone** (95%CI; *p*-value)	**Difference between recall rates for different age groups** (95%CI; *p*-value)
Retrospective, American studies set in community-based radiology practices with annual screening					
Rafferty et al., 2017	FFDM + DBT: 84243 FFDM: 131,996	Review of screening performance metrics	Age 40-49: 115/1000 Age 50-59: 89/1000 Age 60-69: 77/1000 Age 70+: 70/1000	Age 40-49: 137/1000 Age 50-59: 102/1000 Age 60-69: 89/1000 Age 70+: 78/1000	-22 (*p*<.001) -13 (*p*<.001) -12 (*p*<.001) -8 (*p*<.001)
Conant et al., 2016	FFDM + DBT: 9265 FFDM: 35320	Retrospective analysis of data from three PROSPR consortium sites (NB mammography system used not stated)	Age 40-49: 11.4% Age 50-74: 7.3% (*p*<.0001)	Age 40-49: 14.8% Age 50-74: 8.9% (*p*<.0001)	-
Sharpe et al., 2016	FFDM + DBT: 5703 FFDM: 80,149	Prospective study with a retrospective cohort performed at a single site using Hologic Dimensions system for DBT and GE Senographe Essential, 2000D and DS systems	Overall: 6.10 Age 40-49: 8.66%% Age 50-59: 6.31% Age 60-69: 3.66% Age 70+: 4.66%	Overall: 7.51 Age 40-49: 10.93% Age 50-59: 7.20% Age 60-69: 5.86% Age 70+: 5.64%	18.68% (*p*<.0001) 20.78% (*p*=.0075) 12.35% (p=.1657) 37.50 (*p*=.0006) 17.46% (*p*=.2375)

Study	Study participants with dense breasts (BIRADS 3 or 4)	Study type	**FFDM + DBT** (95%CI; *p*-value)	**FFDM alone** (95%CI; *p*-value)	**Difference between recall rates for different age groups** (95%CI; *p*-value)
McDonald et al., 2015	FFDM + DBT: 33,740 FFDM: 10,728 *12079 had one screen* *6293 had two screens* *3023 had three screens*	Retrospective review of mammography metrics from a single site over four years of screening with single reading using Hologic Dimensions system	Overall: 16% Younger than 50: 16.1% 50 or older 16%	Overall: 20.5% Younger than 50: 21.2% 50 or older: 19.5%	22% decrease (*p*=.002) 24% decrease (*p*=.005) 17.9% decrease (*p*=.12)
McCarthy et al., 2014	FFDM + DBT: 5056 FFDM: 3489	Observational study using Hologic Dimensions system for women presenting for mammographic screening at a single institution	Younger than 50: 12.3% 50 or older: 7.3%	Younger than 50: 14% 50 or older 8.8%	12% decrease (*p*=.02) 17% decrease (*p*<.001)
Haas et al., 2013	FFDM + DBT: 6100 FFDM alone: 7058	Retrospective analysis using Hologic Selenia and Dimensions systems	Age 40-49: 10.4% Age 50-59: 7.6% Age 60-69: 7.4% Age 70+: 6.7%	Age 40-49: 16.3%% Age 50-59: 10.6%% Age 60-69: 10.7% Age 70+: 7.9%%	35.8% decrease (*p*<.01) 28% decrease (*p*<.01) 30.3% decrease (*p*=.01)) 15.4% decrease (p=.38) (*not statistically significant)*
Rose et al., 2013	FFDM + DBT: 4666 FFDM: 7009	Observational study following practice implementation of DBT	<50 years: 6.5% 50-64: 75.1% >64: 74.2%	<50 years: 10.3% 50-64: 7.6% >64: 7.9%	37.2% decrease 32.9% decrease 46.6% decrease

3.4. Radiation dose

DBT and mammography are radiation-emitting procedures. Dose is cumulative. The radiation dose required to gain an accurate image is calculated using breast characteristics such as thickness and glandular composition. Balancing safe radiation dose with adequacy of dose needed to acquire clear images is a concern with all such procedures. As Haas et al. (2013) noted

> *"the lifetime attributable risk of a radiation-induced fatal cancer from a single digital breast tomosynthesis acquisition performed at age 40 years is 1.3-2.6 cancers per 100,000 examinations."*

Therefore, an important concern in deciding whether to implement DBT into population-based screening is whether the benefits of the technology (increased cancer detection, reduced recall rates and false positive recall rates) outweigh the increased risk from lifetime exposure to radiation for women participating in breast cancer screening.

In breast imaging, the risk indicator for radiation dose is the MGD[16]. This is the average dose absorbed during image acquisition. In 2013, Sechopoulos published a review of the DBT image acquisition process, including a description of how the MGD is calculated. To estimate MGD, Monte Carlo methods that simulate the acquisition process were used to quantify the energy that the glandular portion of the breast is exposed to. MGD is measured in milligray (mGy).

[16] MGD apportions dose to the at-risk fibroglandular breast tissue (Vedantham et al., 2015).

Key findings

Radiation dose varies with image acquisition process used (DBT or FFDM or combination mode), the number of and type of views, the use of automatic exposure control, breast size and composition, and by DBT system used.

Two-view DBT (i.e., DBT_{MLO} + DBT_{CC}) results in a similar radiation dose compared with FFDM. Almost all the studies included in this literature review assess a screening strategy based on FFDM + DBT compared to FFDM alone (i.e., they integrate digital mammography and DBT together) but the literature on radiation dose also explores different view combinations. FFDM + DBT results in the highest MGD compared to FFDM alone. Other possible combinations including DBT_{MLO} + DM_{CC} result in lower rates. Regardless, the dose associated with FFDM + DBT is still lower than overall dose limits set by international agencies.

More recent studies have investigated the efficacy of DBT + s2DM, which eliminates the need for a separate 2D image acquisition. Using this approach, 2D mammography images are synthesised from a 3D DBT-acquired dataset. This approach halves the effective dose of combined FFDM + DBT, making it comparable to FFDM alone.

Moving to FFDM + DBT as the preferred screening strategy could have significant implications for cumulative dose if separate acquisitions are used for 2D and 3D images, if the screening interval is annual rather than biennial, or if women start participating in mammography-based breast cancer screening in their early 40s.

In this literature review, 17 articles (generated from 11 studies) reported on MGD and radiation dose.

Systematic and/or literature reviews

Four reviews: Houssami (2017); Coop et al., (2016); Svahn et al., (2015A); Vedantham et al., (2015)

RCTs

None

Prospective studies

Two studies: Lång et al., (2016A); Skaane et al., (2013B)

Retrospective studies (observer performance or single-site analysis)

Nine studies: Aujero et al., (2017); Rodriguez-Ruiz et al., (2017); Durand et al., (2015); Shin et al., (2015); Zuley et al., (2014); Sechopoulos (2013); Gur et al., (2012); Olgar et al., (2012); Zuley et al., (2010)

Practical or technical evaluations

Two papers: Strudley et al., (2015); Strudley et al., (2014)

3.4.1. Technical data about radiation dose: Hologic's Selenia Dimensions systems

The first FDA-approved, and the DBT system most commonly used in the literature, is Hologic's Selenia Dimensions DBT system. Almost all studies included in this literature review used this system to acquire and interpret images. For this reason, and because this literature review is not a comparative assessment of system performance, we have presented detailed technical data

about the Dimensions system only. Where this system has not been used in a study about radiation dose (or other screening metric) study, that has been noted in the analysis. Care is needed when extrapolating the findings of this literature review to other DBT systems as the number of images acquired per compression and the time needed to acquire these differ (which may affect the radiation dose).

Figure 2 (overleaf), reproduced from the NHS Hologic Technical Evaluation (Strudley et al., 2014), shows the MGD for FFDM and DBT exposures under automatic exposure control. Information on view is not provided. MGD increases with equivalent breast thickness (which is a factor of both compression and breast size). Women with larger breasts can expect to receive a larger radiation dose to ensure that images acquired are of an acceptable quality. While DBT requires a higher MGD compared to digital mammography per view (and the dose required increases with equivalent breast thickness), it is still below the radiation dose limit for both image acquisition processes.

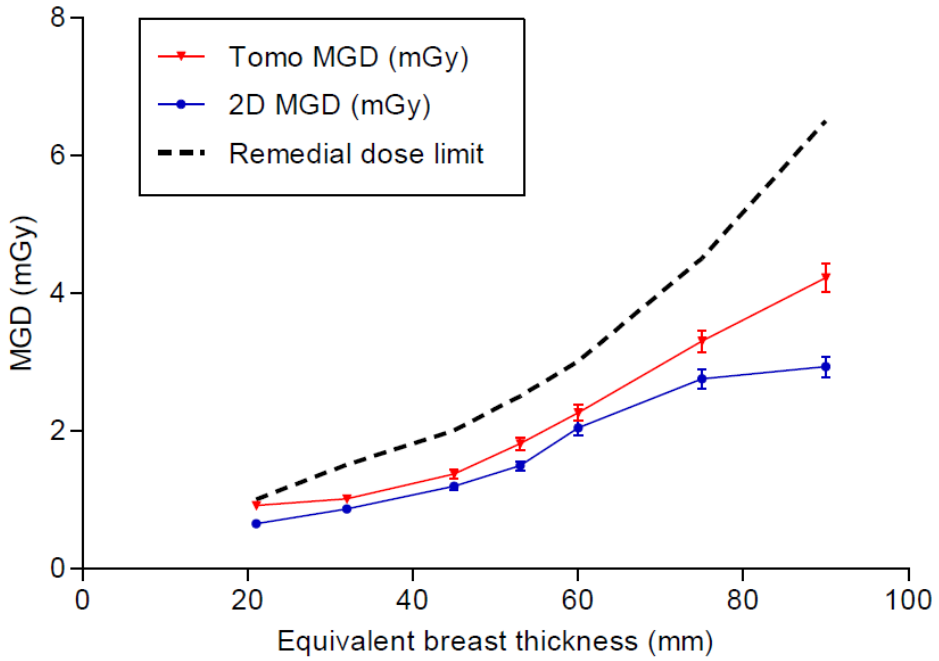


Figure 2: MGD for 2D and DBT exposures under Automatic Exposure Control for Hologic Dimensions system (reproduced from Strudley et al., 2014)

Using a definition of an average breast (53mm thick), Strudley et al. (2014) reported an MGD of 1.81mGy for Hologic's Selenia Dimensions system compared to 1.49mGy in 2D mode. Sechopoulos (2013) found that the DBT acquisition with the Hologic system resulted in an MGD 8 percentage points higher than FFDM alone. However, more recently, a new breast definition has been proposed to better represent the average breast (i.e., the average breast is now 60mm thick with a 14.3% glandular fraction). For these measurements, Sechopoulos found a larger dose increase of 83% from FFDM to DBT (information on view not provided). The author noted that these findings were system-specific to the Hologic Dimensions system. MGD for Siemens Mammomat is comparable: 1.99mGy compared to 0.99mGy in 2D mode (Strudley et al., 2015).

Figure 3 (overleaf) (reproduced from Vedantham et al., 2015) describes the relationship between MGD from one-view (CC-equivalent) for DBT, FFDM and FFDM + DBT compared to compressed breast thickness. It demonstrates that MGD increases with increased compressed breast thickness and shows that for the same compressed breast thickness, the ratio of the MGD

from DBT compared to that from FFDM demonstrates a decreasing trend with increasing glandular fraction. MGD from FFDM and DBT for a single view increased with increasing breast thickness. Compressed breast thickness may be impacted by such things as breast size and level of compression applied. As noted in Chapter 5 of this report, women reported better comfort with lower compression; however, lower compression must be balanced with attaining an acceptable image quality and MGD.

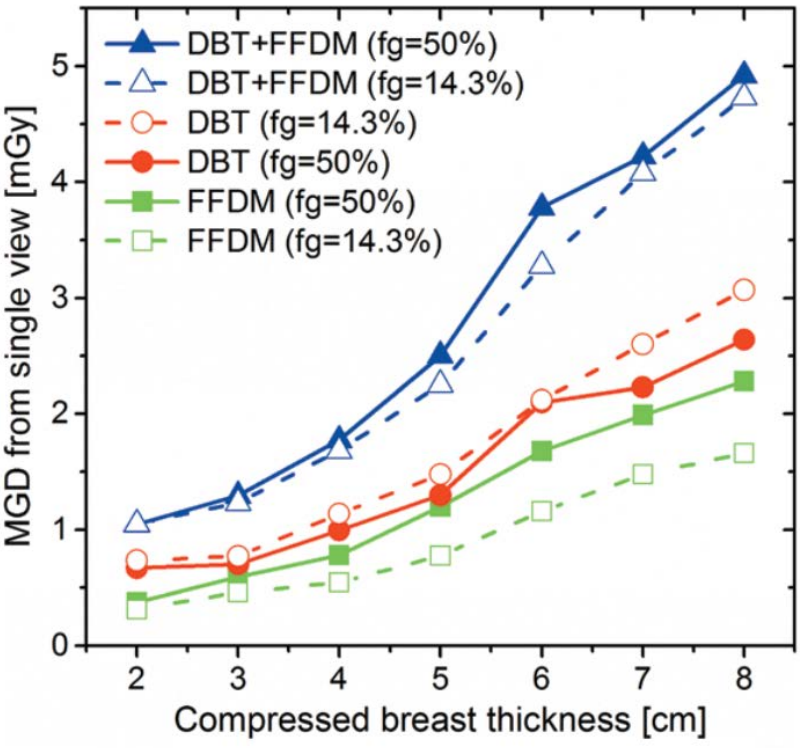


Figure 3: MGD for a single CC-equivalent view from FFDM, DBT, and the combined DBT-FFDM provided for various compressed breast thicknesses and for 50% and 14.3% fibroglandular breasts (reproduced from Vedantham et al., 2015)

Maximum acceptable limit for MGD

The American College of Radiology Mammography Quality Standards (updated in 2004) prescribed a radiation dose limit of 3.0mGy per breast per CC (equivalent) view (Mammography Quality Standards Act Regulations, Part 900.12(e)(5)(vi), cited in Aujero et al., 2017). This is higher than the MGD dose limit set for the BreastScreen Australia program (which is 2.0mGy for a 50% adipose/50% glandular breast) (BreastScreen Australia, 2015). Based on data from the NHS technical evaluation (Strudley et al., 2014), MGD for both DBT alone and FFDM are below the remedial dose limit curve. Regardless, radiation dose always remains a concern, and the intention of all screening programs is always to reduce radiation exposure as much as possible.

3.4.2. Results on comparative radiation dose

MGD can be affected by automatic exposure control, image quality required, patient positioning (particularly to get a clear MLO view), breast composition, size and mammography/DBT system (Vedantham et al., 2015). Such variations may underpin the performance differences reported in the literature, which are based on real-world studies. These 'real-world' studies are discussed below.

Results for radiation dose for FFDM + DBT compared to FFDM alone

Fifteen studies reported radiation dose comparing FFDM + DBT to FFDM alone.

Systematic reviews and literature reviews

One systematic review (Coop et al., 2016) discussed the radiation dose of FFDM + DBT compared with FFDM alone, largely based on data from the STORM and OTS trials. Coop et al.'s systematic review (including information about study inclusions/exclusions) is described in *section 3.1*. At the time of image acquisition in the studies reviewed by Coop et al., Hologic systems were the only FDA-approved DBT units and were the systems that all studies included in the systematic review were based on. Coop et al. did not perform a pooled analysis of radiation dose.

Coop et al. reported that radiation dose from DBT alone could be up to 30 percent lower than FFDM alone based on results from a Monte Carlo phantom study (citing Baptista et al.'s 2014 work). Coop et al. did not discuss differences in MGD by view (or combination of view). Therefore, we have presumed that their analysis reported results are for two-view DBT and FFDM.

Study Design Quality	Results: MGD per view
Skaane et al., 2013B OTS trial High quality	FFDM + DBT: 3.53 mGy FFDM alone: 1.58 mGy Average breast thickness: 54mm

Because Coop et al. did not provide detailed discussion of the MGD reported for the OTS trials, but these results are reported in the primary papers, we have reported these rates in the table (above). Skaane et al. (2013B) reported that the dose for FFDM + two-view DBT was 2.24 times higher than that reported for FFDM alone (an increase of 1.95 mGy). Both images were acquired using a single compression. In the OTS trial, when used in two-view combination mode, this is slightly over double the radiation dose compared to FFDM alone.

Coop et al. also cited four smaller studies, which noted that the radiation dose for DBT and FFDM alone is approximately the same, and therefore the radiation dose for FFDM + DBT is about double that of FFDM alone. All the studies included in Coop et al.'s review stated that radiation doses (measured as MGD) were below the maximum per view set out by the FDA of 3.0mGy per acquisition and were below FDA-approved limits for acceptable risk.

Svahn et al.'s (2015A) literature review reported on radiation dose ratios presented in 17 papers comparing radiation dose estimates for DBT to FFDM. Different studies used different views and combinations of views (see bullet points below). Pooled analysis was not performed. Five different DBT systems were used (GE, Hologic, Siemens, XCounter and Sectra), although only the Hologic system had FDA approval (the other models were prototypes). Svahn et al. found that, using the Hologic system for DBT and other systems for FFDM, DBT was associated with a wide range of doses depending on the view and whether DBT was integrated with FFDM or used alone:

- In seven studies using DBT_{MLO}, dose ratios (D_{DBT}/D_{FFDM}) ranged from 0.34 to 1.0 (that is, the radiation dose is lower than or comparable to FFDM).
- In five studies using two-view DBT, dose ratios (D_{DBT}/D_{FFDM}) ranged from 0.68 to 1.7 (i.e., providing a lower dose or up to 70% more than FFDM alone).

- In studies combining DBT_{MLO} with FFDM, radiation dose ratios increased slightly: 1.03 to 1.50 (*NB* higher dose ratio was associated with FFDM; lower dose ratio was associated with DM_{CC}).
- For FFDM + two-view DBT, the dose ratio (D_{DBT}/D_{FFDM}) was slightly more than double: 2.0 to 2.23.
- For two-view DBT + s2DM, dose was reduced.

For combined FFDM + DBT, radiation doses were elevated and like those reported in Coop et al. Svahn et al.'s findings reported that when two-view DBT alone is used, it results in a generally similar dose as FFDM alone. For FFDM + DBT, the dose levels were substantially higher (about twice that of FFDM alone). Svahn et al. also discussed the use of s2DM, which is discussed below.

Randomised controlled trials and prospective studies

No information about radiation dose was provided in Maxwell et al.'s 2016 RCT. Information from the STORM and OTS trials is discussed in Coop et al.'s systematic review.

Retrospective observational studies

This review identified four retrospective observational studies that discussed radiation dose (Durand et al., 2015; Shin et al., 2015; Olgar et al., 2012; Zuley et al., 2010).

The following results mirror the findings from the Hologic Selenia Dimensions technical evaluation (Strudley et al., 2014), and Svahn et al.'s review that the MGD varies depending on which combination and views are used to acquire images. Overall, MGD for all combinations studied to date fall below approved quality standards limits.

Shin et al. (2015) conducted a retrospective performance study of 149 women to compare MGD (reported as average glandular dose) between DBT_{MLO} plus DM_{CC} with FFDM. While the study population included some symptomatic women, Shin et al.'s study provides useful information about radiation dose and has therefore been included. Shin et al. found that the mean MGD of DBT_{MLO} + DM_{CC} was significantly higher than that of FFDM (p<.001). Breast thickness was significantly associated with the MGD of the screening strategies. Both MGDs increased with breast thickness (p<.001). Shin et al. also found that MGD was affected by density as well, reporting that MGD increased in women with dense and thick breasts compared to fatty and thin breasts. Finally, they concluded that single view DBT plus single view 2D digital mammography decreased radiation dose and improved diagnostic performance (reflecting that the risks associated with radiation exposure need to be balanced with improved detection and diagnosis) compared to FFDM. Shin et al. also noted that for women with dense breasts (BIRADS levels 3 to 4), sensitivity was increased to 89.8% with DBT_{MLO} + DM_{CC} compared to FFDM (83.1%) although this was only slightly higher than the increase in sensitivity reported for women with less dense breasts (88.2% with DBT_{MLO} plus DM_{CC} compared to 78.4% for FFDM).

Durand et al. (2015) compared the radiation dose for one DBT view, one 2D view and one DBT + 2D digital mammography view but did not specify which projections (MLO or CC) were used. The authors reported that the radiation dose for a one view FFDM + DBT examination was higher than that for one 2D view. One view DBT also had a high radiation dose compared to 2D alone. Overall, the doses for all views were still below the FDA radiation dose safety limits. Durand et al. reported lower doses for all screening strategies than the other studies. This may be because the average breast thickness of that study is lower than that for other studies. Higher breast thickness requires a higher radiation dose to gain effective images.

Zuley et al. (2010) reported that the radiation used to acquire DBT images sets was the same as that used to acquire standard FFDM images, with average mid-breast dose of about 2mGy per view. No further specification of radiation dose was provided (therefore this study is not included in Table 25, overleaf).

Table 25: MGD reported in retrospective studies comparing FFDM + DBT to DBT (either 1 or 2 view)

Study	Sample	System used	Average breast thickness (mm)	FFDM + DBT MGD (mGy) per view	DBT (2-v) MGD (mGy) per view (SD)	DBT (1-v) MGD (mGy) per view	FFDM (1-v) MGD (mGy) per view	FFDM (2-v) MGD (mGy) per view (SD)
Durand et al. 2015	FFDM + DBT: 8591 FFDM: 9364	Bilateral DBT and FFDM views acquired under a single compression using Hologic Selenia Dimensions system	42 (*NB breast phantom)*	2.65		1.45	1.2	
Shin et al. 2015	149 Korean women (including 61 asymptomatic women recalled from screening) referred for diagnostic work up Median age = 50y	Bilateral FFDM + DBT using Hologic Selenia Dimensions system.	48	NR	1.74 (0.93 – 5.02) (*p*<.001 compared to MGD for FFDM)			1.63 (0.68 – 7.41)

Olgar et al. (2012) conducted a retrospective study of 2247 conventional FFDM images and 9845 DBT images from 641 women examined with Hologic Selenia Dimensions to determine the MGD for each view and each screening strategy. The authors concluded that the MGD per exposure for DBT was on average 34% higher than for FFDM for women examined with the same compressed breast thickness, with MLO views requiring a higher dose to obtain the image compared to CC views. This is slightly higher than the results of the NHS technical evaluation (Strudley et al., 2014) although the compressed breast thickness used in Olgar et al.'s study was larger, which could account for this difference.

Table 26: MGD results from Olgar et al. (2012) for CC and MLO views

View type	Compressed breast thickness (mm)		DBT alone (mGy)	FFDM alone (mGy)
MLO view	56.0	MGD per exposure for the standard breast	2.29	1.66
		MGD after correction for real breast composition	2.63	1.94
CC view	52.7	MGD per exposure for the standard breast	2.19	1.57
		MGD after correction for real breast composition	2.53	1.82

Results for radiation dose for DBT + s2DM compared to FFDM + DBT or FFDM alone

A software release for the Hologic system (C-View) has enabled a combination mode which does not require an actual additional radiation exposure to acquire 2D images. Instead, 2D images are generated from the DBT-acquired data. Eliminating the 2D exposure shortens the acquisition and compression time and reduces the radiation dose for FFDM + DBT by about half (from 3.3mGy to 1.81mGy) (Strudley et al., 2014).

Five studies reported 'real-world' radiation dose comparing DBT + s2DM to FFDM + DBT or FFDM alone. Three were described in Houssami's 2017 literature review and two others (Zuley et al., 2014; Gur et al., 2012) reported separately.

Systematic reviews, RCTs and literature reviews

No systematic reviews or RCTs compared MGD from DBT + s2DM to FFDM + DBT or FFDM alone.

In August 2017, Houssami (2017) conducted a literature review of population-based breast cancer screening clinical outcomes and performance metrics (including radiation dose) for the following screening strategies: DBT + s2DM compared to FFDM + DBT or FFDM alone. The literature review reported on three articles that discussed radiation dose from trials including two reporting on the STORM-2 and OTS trials (Bernardi et al., 2016; Skaane et al., 2014). Using information from the prospective trials, Houssami reported a substantially lower MGD for DBT + s2DM: 55% to 58% of the MGD per view for FFDM + DBT (specific data on the DBT + s2DM is provided in the DBT alone column of Table 27, below). Houssami reported that the ability to synthesise 2D images of the breast from the DBT acquisitions means that the need for dual acquisitions (DBT and FFDM) can be avoided, which could alleviate some of the concerns about double radiation dosing to the breast for routine screening (as discussed in the previous section). Zuckerman et al. (2016) also reported a total dose that was 39% lower for DBT + s2DM than FFDM + DBT (p<.001). Neither Zuckerman et al. nor Houssami provided an explanation for why the dosages in Zuckerman et al. were higher than those of the other studies.

Table 27: Studies included in Houssami's 2017 literature review

Study	**Sample**	**System used**	**Average breast thickness** (mm)	**FFDM + DBT** MGD per view in mGy (SD)	**DBT alone** MGD per view in mGy (SD)	**FFDM alone** MGD per view in mGy (SD)
Prospective fully paired trials embedded in European population-based screening programs with biennial screening						
Bernardi et al., 2016 (STORM-2)	9672 asymptomatic Italian women aged 49 years or older (median age 58 years) who attended population-based screening	Selenia Dimensions system with C-view for FFDM + DBT with single reading	NR	3.22 (1.16)	1.87 (0.67)	1.36 (0.51)
Skaane et al., 2014 OTS trial	12,270 screens from 24,901 Norwegian women aged 50-69 years (mean age 59.2 years)	Selenia Dimensions system with C-view for FFDM + DBT with double reading and two study periods (one using C-view, one using an earlier software. Only rates using C-view are reported here)	53.9 (12.8)	3.53	1.95 (0.58)	1.58 (0.61)
Retrospective American studies set in community-based radiology practices with annual screening						
Zuckerman et al., 2016	15,571 American women screened with FFDM + DBT and 5366 women	Bilateral CC and MLO images were obtained for each screening study using Hologic Dimensions system	60.8	7.97 (p<.001)	4.2-4.88	3.77

Study	Sample	System used	Average breast thickness (mm)	FFDM + DBT MGD per view in mGy (SD)	DBT alone MGD per view in mGy (SD)	FFDM alone MGD per view in mGy (SD)
	screened with DBT + s2DM					

Retrospective observational studies

Gur et al. (2012) conducted a retrospective laboratory study on a set of FFDM and DBT images performed on 118 women using the Hologic system. Images were acquired with a combination screening strategy, with FFDM acquired first followed by DBT. The radiation dose used to acquire the DBT images was approximately the same as that for the FFDM acquisition (about 2mGy per view). Reconstruction of synthetic 2D images from the 3D data set did not require any additional radiation exposure. Therefore, the authors concluded that acquisition of DBT + s2DM information uses the same radiation exposure as FFDM alone, and about half that of FFDM + DBT. The reconstructed images were of a reasonably high quality and acceptable for the detection of cancer. The difference in the results from this study compared to those discussed in Houssami's literature review may result from improvements to the software algorithms and reconstruction (given that C-view was made available after Gur et al.'s study).

Another study by Zuley et al. (2014) assessed radiation dose for DBT + s2DM versus FFDM alone and FFDM + DBT. The authors found that DBT + s2DM delivered a lower dose than that required for FFDM + DBT, but again did not provide any specific dose measurements.

s2DM is a promising alternative to FFDM + DBT in terms of reducing the necessary radiation dose, and therefore reducing the harm to women that is potentially caused by population-based screening. However, more work needs to be done to ensure that the quality of images produced by s2DM is of the same standard as acquired digital images, particularly in terms of classifying microcalcifications (see *section 3.1.3*).

Radiation dose results for DBT_{MLO} compared to other imaging combinations

In the active Malmö trial (Lång et al., 2016A), image acquisition consists of FFDM immediately followed by DBT_{MLO}. In contrast to the other studies cited in this review, the Malmö trial uses Siemens' Mammomat Inspiration system. Lång et al. (2016A) reported that the absorbed radiation dose in a DBT_{MLO} examination is approximately 70% of the absorbed dose in FFDM. The FFDM + DBT combination mode used in other screening trials gives additional radiation dose compared to the single screening strategy used in this study (Lång et al., 2016A). The authors also noted that the use of s2DM could provide a means to sustain low radiation doses when using combination modes.

Study Design Quality	Results: MGD per view
Rodriguez-Ruiz et al., 2017	$DBT_{MLO:}$ 2.41 mGy DBT_{MLO} + DM_{CC}: 3.62 mGy FFDM: 2.41 mGy FFDM + DBT: 7.23 mGy
Lång et al., 2016A Trial (Malmö) High quality	FFDM alone: 1.2 mGy Two-view DBT alone: 1.6 mGy Average breast thickness: 53mm

Rodriguez-Ruiz et al. (2017) compared DBT_{MLO} to DBT_{MLO} + DM_{CC}, FFDM and FFDM + DBT using Siemens Mammomat Inspirations DBT system. The authors reported variances in MGD by the number of views taken, all of which were much higher than the results reported from the Malmö trial (and they used the same system). Another reported difference between the studies was that DBT_{MLO} MGD was the same as FFDM. Comparisons of MGD for other view combinations were all higher compared to the MGD for FFDM and for rates reported on Hologic's system. It is not clear whether these differences are due to the Mammomat system or the scan angle (which was wider in Rodriguez-Ruiz's study, whereas the Malmö study used a narrow angle system), or both.

4. IMPLEMENTATION OF DBT AS A SCREENING TOOL

Much of the published literature focuses on DBT's sensitivity, specificity and safety with the studies on the nature of the association between DBT (either alone, as an adjunct to FFDM, or with s2DM) and specific clinical outcomes and performance metrics. There is growing confidence that as a screening strategy DBT could enhance a screening program (although significant research gaps relating to long term mortality benefit remain). Another key area of research to consider is implementation. That is, if DBT was to become a preferred screening strategy, what issues need to be considered to ensure the maximum benefits accrue to women and health practitioners. Key issues to consider are:

- Image acquisition
- Reader performance: experience and accuracy
- Interpretation time requirements
- Other interpretation considerations, and
- Cost.

Key findings

Several issues need to be considered to ensure the maximum benefits of DBT are realised before it is implemented as a preferred screening strategy.

DBT images requires a small amount of additional time per view to acquire images. Additionally, interpretation times have been shown to increase to varying degrees, but the interpretation times tend to decrease as readers gain experience in interpreting DBT.

Detailed analysis of the cost effectiveness has not been performed in jurisdictions other than the United States. American modelled analyses show that DBT is cost-effective (in terms of finance) for community-based practices (although these are very specific to the insurance programs).

As discussed at *section 1.4,* the PROSPECTS trial and the Maroondah trial will further investigate the cost effectiveness of breast cancer screening using FFDM + DBT compared to FFDM or DBT + s2DM. The PROSPECTS trial will be conducted over seven years from 2018, with initial results to be presented within 18-24 months. Maroondah will report earlier than this date. Both trials will provide useful information about implementation considerations which will be useful advice for national and state screening programs as well.

In this literature review, 34 studies reported on these implementation issues.

Systematic and/or literature reviews

Five studies: Houssami (2017); Poplack, (2017); Coop et al., (2016); Gilbert et al., (2016); Houssami and Skaane (2013)

Practical evaluations

One study: Mungutroy et al., (2014)

RCTs

One study: Maxwell et al., (2017)

Modelled analyses

Four studies: Miller et al., (2017); Kalra et al., (2016); Bonafede et al., (2015); Lee et al., (2014)

Prospective studies

Eight studies: Hunter et al., (2017); Carbonaro et al., (2016); Lång et al., (2016A); Lång et al., (2016B); Bernardi et al., (2014); Ciatto et al., (2013); Skaane et al., (2013); Bernardi et al., (2012)

Retrospective studies (observer performance or single-site analysis)

Fifteen studies: Rodriguez-Ruiz et al., (2017); Sharpe et al., (2016); Durand et al., (2015); Dang et al., (2014); Rose et al., (2014); Haas et al., (2013); Lee et al., (2014); Rafferty et al., (2013); Zuley et al., (2013); Bernardi et al., (2012); Gur et al., (2012); Wallis et al., (2012); Svane et al., (2011); Gennaro et al., (2010); Svahn et al., (2010)

4.1. Image acquisition

Acquisition of DBT images requires an additional time per view to acquire the images, ranging from about 10 seconds to almost a minute depending on the projection, number of views and the DBT system used.

The NHS practical evaluation (Mungutroy et al., 2014) determined the timing of DBT examinations taken using Hologic Dimensions systems by examining the start time of the whole examination and the start time of each individual exposure. The length of each exposure is 1.2 seconds for a 2D exposure of an average breast, and 6 seconds in total for a series of DBT exposures. In total, the time for two combined views would be about 2 minutes 49 seconds. The time taken to acquire the 3D project images and the conventional 2D image is not significantly longer than the time required to perform FFDM alone (Hardesty, 2015). Each view is obtained during the same breast compression as standard digital mammographic projections. There is no need for the woman to be repositioned during the examination (except for the repositioning already associated with moving from acquiring the CC image to the MLO image). Therefore, it is usually associated with only a small amount of extra time investment for women and technologists (Mungutroy et al., 2014).

In a 'real world' screening program (STORM), Bernardi et al. (2012) reported on the average acquisition time measured from the start of the first view of the breast positioning to compression release at the completion of the last view. Average acquisition time for seven radiographers, based on 20 screening examinations, was longer for FFDM + two-view DBT (4 min 3 s; range 3min 53 s– 4min 18 s) than FFDM alone (3 min 13 s; range 3min 0 s–3 min 26 s; p=.01). The 'real world' acquisition process is much longer for both FFDM alone and FFDM + DBT than the practical evaluation data from the NHSBSP. This may be due to familiarisation with the technique.

Reporting on the OTS trial, Skaane et al. (2013) obtained two views (CC and MLO) of each breast with FFDM and DBT with single breast compression per view. DBT images required about 10 additional seconds per view to obtain (an additional 40 s overall with the use of DBT as an adjunct screen; 3 min 55 sec), which is like that reported from STORM.

4.2. Reader performance: experience and accuracy

Most studies involve readers with a range of experience (from relatively newly trained to much more experienced radiologists). There was also variation by level of experience in interpreting breast images or more general radiology, with studies reporting that less experienced radiologists improve their accuracy more than more experienced radiologists. It is not clear whether this improvement merely reflects the development of a less experienced practitioners' professional competence, or whether it reflects that DBT is 'easier' to read without as much experience in breast cancer imaging.

Results for FFDM + DBT

Coop et al. (2016) discussed whether DBT aids less experienced radiologists in interpreting mammography and cited two studies which discussed reader performance: the STORM trial (Bernardi et al., 2014) and another smaller retrospective trial. They found that less experienced radiologists improved their accuracy and increased rates of detection and specificity with the use of DBT. For example, seven of eight radiologists improved individual detection. Results ranged from 0% improvement in detection to 54% improvement (Bernardi et al., 2014). In another study, the use of DBT resulted in improved performance (as measured by increased detection rates) of between 30 and 60% (Rose et al. 2014).

Data from the Malmö trial (Lång et al., 2016A) also reported that inexperienced radiologists had a higher specificity and lower sensitivity for DBT_{MLO} compared to FFDM. Six readers participated in this trial: five readers with more than 10 years of experience in breast radiology and one reader with less than 10 years (average of 26 years of experience, with a range of 8 to 41 years). It was hypothesized that less experienced radiologists could become as accurate as the more experienced readers following further training. It was also observed that single view DBT took an average of 25 percent longer to read than FFDM. It was noted that additional training of radiologists on reading DBT might speed up interpreting times.

Coop et al. noted that even though DBT has better sensitivity than mammography, inter-observer variability of readers must be addressed (Rose et al., 2014). Inter-reader performance variability is discussed below.

In 2016, Carbonaro et al. reported on a fully paired prospective trial conducted within a screening program of 280 women. It found that variability in recall rates and detection was higher when reading digital mammography alone, that is, that FFDM + FFDM improved the agreement between readers. The authors found a two-fold increase in inter-reader agreement when DBT was used in conjunction with FFDM. The authors also noted that interobserver agreement for DBT + s2DM was higher than those for FFDM. Sharpe et al. (2016) found variations in the radiologist-specific recall rates for both 2D and DBT interpretations. They noted that since DBT is a new screening technology, there is likely to be a learning curve that will occur at different rates for different radiologists.

Reader uncertainty as well as reader variability has also been investigated. Maxwell et al. (2017) demonstrated that increased reader uncertainty was caused by DBT in the first screening round (although this was reversed in the second round). The authors noted that this was despite all readers being trained in DBT and having experience using it routinely in screening assessment before the study began. Maxwell et al. hypothesised that uncertainty may be related to the need to develop confidence in dismissing asymmetric densities on DBT appearance alone. These results suggest that a learning curve exists, with reader uncertainty reversing with increased experience.

Rafferty et al. (2013) expressed some concerns about inappropriate dismissals of masses by some of their readers in the first of their reader studies. In reader study 1, cancers manifesting as certain types (particularly circumscribed lobulated masses), were being inappropriately dismissed by some readers. The authors noted that in mammographic interpretation, radiologists often associate circumscribed masses with a benign process. However, in DBT imaging, circumscribed margins may be an indication of malignancy. The authors recommended the importance of emphasising this point when transitioning to DBT screening. Training for the second reader study reinforced these principles, and nearly all the readers correctly classified circumscribed, lobulated lesions. These results emphasise the importance of effective training in the clinical environment to avoid inappropriate dismissal of some cancers, particularly when introducing new screening methods such as DBT.

Data from the Malmö trial (Lång et al., 2016A) found that when using DBT, different reading strategies could impact on overall accuracy. The authors reported that DBT_{MLO} was sensitive and specific enough, and extra views did not result in further increases in incremental CDR. The authors estimated that reading time for DBT_{MLO} in screening would be doubled compared to FFDM (but the actual reading time was not registered in the study). This is the only study to assess reading with different single views, and further research is needed to reproduce Lång et al.'s findings. Rosso et al. (2015), also reporting on Malmö trial data, commented that there appeared to be more limited differences in reader performance by experience (that is, no differences in improvement were noticed).

4.3. Interpretation time requirements

The literature reports a strong, consistent theme that implementation of DBT as an adjunct to FFDM increases interpretation time. DBT produces many more images than FFDM alone (up to 25 compared to two). As such, readers need to look through more images to complete the reading and interpretation of screening results. All studies reporting on reading time reported that reading time is increased (usually by double) although no studies reported on reading DBT images only; they all reported reading times associated with DBT as an adjunct screen to FFDM. Increased reading time will have workflow and reader/radiologist resourcing implications.

Poplack (2017) reviewed three prospective population-based studies (the STORM, OTS and Malmö trials), and concluded that the "*only cautionary outcome*" for FFDM + DBT was the time required for image interpretation, which was about double that of FFDM alone (approximately 91 seconds compared to 45 seconds). Gilbert et al. (2016) also reported on four studies (including the STORM and OTS trials and two retrospective analyses) where reader time was approximately double for DBT compared with FFDM. These results are described in Table 28 (overleaf).

Table 28: Reader time reported in Gilbert et al. (2016)

Study	Sample	Study type	Reading time
Prospective trials embedded in European population-based screening programs with biennial screening			
Ciatto et al., 2013 (STORM) *Main article*	7292 asymptomatic, average risk women	Prospective, fully paired trial using Hologic Selenia Dimensions systems in combination mode	Reading time doubled
Skaane et al. 2013 (OTS trial)	12,631 women aged 50 -69 years participating in the biennial Oslo breast screening program, with nine months follow-up.	Prospective, fully paired trial using Hologic Dimensions system with double reading	Reading time of 45 seconds for FFDM and 91 seconds for DBT
Retrospective, American studies set in community-based radiology practices with annual screening			
Zuley et al. 2013	125 selected examinations, 35 with verified cancers and 90 negative for cancer	Twice interpreted using FFDM alone followed by a combined FFDM + DBT mode.	Increase in reading time of 33%
Wallis et al. 2012	10 readers classified 130 cases	FFDM compared with DBT_{MLO+CC} and DBT_{CC} using a multi-reader, multi-case receiver characteristic method	67 seconds for FFDM and 124 seconds for DBT

In 2014, Dang et al. conducted a study on the effect of adding DBT to FFDM on image interpretation time. Ten radiologists read and interpreted images from 3665 examinations (1502 FFDM + DBT and 2163 FFDM images). Of the radiologists, two had more than 20 years' experience, four had 10-15 years, three had 5-10 years, and one had fewer than five years of breast imaging experience. All radiologists (apart from one very experienced radiologist) took a statistically significant longer time to interpret FFDM + DBT images compared to reading FFDM images. For FFDM + DBT, radiologists interpreted an average of 23.8 screens per hour (an average of 2.8 minutes per screening examination) compared with an average of 34.0 screens per hour for FFDM imaging (an average of 1.9 minutes per screening examination). Therefore, the time taken to interpret a DBT+DM examination was on average 0.9 minute longer (47 percent) than that of FFDM. Dang et al. also found that increased breast imaging experience led to a decrease in the overall additional time required to interpret FFDM + DBT examinations compared with the time taken to interpret FFDM alone examinations. The authors noted that this study would be useful for preparing for the effect of mainstream introduction of FFDM + DBT on radiologists' workload and planning for staffing requirements and resource allocation.

In 2012, Bernardi et al. reported on the incremental effect of DBT on acquisition and reading time. The study found that the average reading time per screening examination was 33 seconds for FFDM alone, and 77s for FFDM + DBT (a statistically significant increase of 135 percent). Bernardi et al. (2012) noted that the FFDM + DBT workstation software is simple and easy to manage, and that the manufacturer (in the US) recommends eight hours of training before managing and reporting 3D images. However, in Bernardi et al.'s study, the training in managing and reporting 3D images was at least ten times longer than the recommended period. From these results, the authors concluded that radiologist/reader workload would be substantially increased with prolonged reading time (on top of the required training time).

Rodriguez-Ruiz et al. (2017) reported similar results to other studies and concluded that additional training of radiologists might speed interpretation times up. For this study, reader times were defined as the time spent on evaluating, scoring and annotating DBT (the first step of the reading session) and FFDM (second reading session). Readers 1, 2 and 3 were relatively inexperienced interpreting DBT compared with readers 4, 5 and 6. On average, it was observed that DBT took about 25% longer to read than FFDM. The authors concluded that additional training of radiologists on reading DBT might enable fast reading of DBT screening strategies.

Reader	DBT reading time in seconds (95%CI)	FFDM reading time in seconds (95%CI)	*p*-value
Reader 1	58 (52–65)	36 (32–40)	<.001
Reader 2	57 (52–63)	51 (46–57)	=.095
Reader 3	42 (37–47)	46 (40–53)	=.281
Reader 4	63 (55–71)	64 (56–71)	=.880
Reader 5	56 (50–62)	31 (27–35)	<.001
Reader 6	55 (48–62)	35 (30–40)	<.001
Average	55 (52–59)	44 (40–48)	<.001

4.4. Other implementation considerations

Lee et al. (2014) reported that DBT has the advantage of increased patient throughput, streamlined equipment needs, reduced physical space needs and reduced training of technical staff. However, the authors expressed concern that these benefits have resulted in many practices adopting DBT at an early stage before the acquisition of sufficient clinical effectiveness data. While such implementation issues are important, ensuring appropriate clinical performance should be a priority for any breast-screening program considering the adoption of DBT.

4.5. Cost of implementing DBT

Most of the available research investigating the cost of implementing DBT focuses on the use of FFDM + DBT compared to DBT alone. Costs associated with the implementation of DBT depend somewhat on existing infrastructure (including whether current FFDM units are capable of a minor software or detector upgrade to become DBT-capable or whether full new systems need to be purchased). Other costs relate to the need for increased capacity for data storage (DBT images are larger and there are more of them compared to FFDM). These extra costs could be offset if DBT results in reduced recall rates and assessment costs associated with assessment screening or biopsy but there is limited literature discussing this. No literature exploring cost was available for the Australian context.

Modelled analyses

This review identified four studies which used models to discuss the cost of implementing DBT as a screening tool (Miller et al., 2017; Kalra et al., 2016; Bonafede et al., 2016; Lee et al., 2014). Table 29 (overleaf) summarises these findings. Most of these studies used modelled analyses to discuss the effect of DBT implementation on insurance programs in the United States and are therefore not directly applicable to the Australian breast screening program. At the date of this literature review (December 2017), no modelling analyses have been conducted elsewhere.

Miller et al. (2017) discussed the effect of DBT implementation on the US Medicaid program. The purpose of the study was to conduct a clinical-economic value analysis of DBT for breast cancer

screening among women enrolled in Medicaid. The model predicted significant total estimated cost savings to the state program, as well as individual per patient savings with the introduction of adjunct DBT to FFDM screening. The results of this study demonstrated that there is potential economic favourability for DBT when considering the clinical benefits of adding DBT to FFDM. However, it should be emphasised that the model used in the study was very specific to the Medicaid program, and it is likely that it does not have wider applicability.

In 2016, Kalra et al. conducted an evaluation of the financial cost-effectiveness of the addition of DBT to FFDM alone in annual screening for women beginning at 40 years old and determined the incremental cost per quality-adjusted life year (QALY) for DBT over FFDM alone for all ages. The study found that DBT achieved higher than expected utility for the overall population as well as all individual age sub-groups. The breakdown is presented in Table 30. The authors found that DBT is most financially cost-effective for women aged 40-49 years.

Patient Age Group (y)	Cost of FFDM (USD)/Effectiveness (QALY)	Cost of DBT (USD)/Effectiveness (QALY)	Incremental net monetary benefit (USD)
All	14,500/15.46	15,312/15.50	3188
40 – 49	4961/8.01	5363/8.03	1598
50 – 59	5043/7.78	5497/7.79	546
60 -69	6401/7.34	6866/7.35	535
>70	7714/6.80	8213/6.81	501

A model by Bonafede et al. (2015) and Lee et al. (2014) similarly demonstrated clinical and economic favourability for DBT in breast cancer screening among commercially-insured US women.

Table 29: Modelled analyses for the cost of implementing DBT

Study	Study type	Economic modelling findings for implementation of DBT
Miller et al. 2017	Modelled analysis of the state Medicaid program	Annual cost savings per patient: USD$8.14 Annual cost savings for a typical state Medicaid program: up to USD$207,000 Annual cost savings nationally: ~ USD$10.7 million
Kalra et al. 2016	Modelled analysis	Total discounted cost of $15,312 and 15.50 QALYs compared with $14,500 and 15.46 QALYs for 2D mammography alone.
Bonafede et al. 2015	Modelled analysis	Annual cost savings per patient: USD$28.53 Annual cost savings national: USD$2.4 million

Cost information from trials embedded in national screening programs

Lång et al. (2016A) stated that the financial cost-effectiveness of DBT compared to digital mammography in screening has yet to be evaluated, and that studies on cost will be important for future decisions on the role of DBT in screening. The authors noted that it is reasonable to assume that DBT will be more expensive than FFDM, but that it is important to relate that to the benefits of increased, and possibly earlier cancer detection. In another paper in 2016, Lång et al. (2016B) stated that the cost of false positive recalls has been estimated to be almost a third of the total cost of a screening program based on FFDM alone. Before implementation of DBT into screening programs, further analyses of the cost-benefit are needed, including Australian-specific data.

Hunter et al. (2017) conducted a retrospective data analysis to determine whether DBT is a financially cost-effective alternative to FFDM for both Medicare and privately insured patients undergoing screening. The study involved data from 6319 women (3655 who underwent DBT and 2664 who underwent FFDM). Private insurance billing cost USD2.9 million, and Medicare cost USD1.2 million for screening, follow-up imaging and radiologic procedures. For the DBT group, per-person costs were about USD40 higher using both forms of insurance. However, per cancer detected, costs were lower for the DBT group for both forms of insurance, leading to a possible USD3.7 million saved per 1000 cancers detected for private insurance and USD899,000 saved per 1000 cancers for Medicare. After standardisation of the difference in cancer detection rates, it was found that DBT was a financially cost-effective alternative to FFDM for private insurance, but not with respect to Medicare. Therefore, the authors concluded that DBT is potentially a financially cost-effective alternative to FFDM.

None of the retrospective studies identified in this review conducted detailed cost analysis of DBT compared with FFDM. However, two studies made general cost statements based on other results. Durand et al. (2015) noted that DBT better assists in the characterisation of benign findings, which decreases false-positive results, and may therefore result in preventing the cost of unnecessary recalls. Haas et al. (2013) also linked decreased costs to reduced recall rates.

Very little detailed cost analysis of screening programs has been conducted to date. However, initial results from one retrospective trial, and results from modelled analyses show that DBT is likely to be a financially cost-effective alternative to FFDM alone in the US. More evidence is needed to determine whether these cost savings will translate to the Australian screening program.

The PROSPECTS trial and the Maroondah pilot will investigate the cost effectiveness of breast cancer screening using FFDM + DBT compared to DBT + s2DM in settings other than the United States. Study results will provide further information about financial cost-effectiveness.

5. ACCEPTABILITY TO WOMEN

We searched for literature that would provide information about women's experiences using and/or attitude towards DBT as a screening tool (either alone or when used in combination with FFDM). We were specifically interested in whether DBT impacted on women's anxiety (either about having a mammogram, participating in a screening program or when dealing with positive mammogram results), women's confidence that DBT would support the early detection of breast cancer and therefore their choice to receive this test, views on discomfort and pain associated with compression, and any issues related to convenience.

Our literature search returned one study that specifically explored the effect of reduced compression used in DBT on pain, anxiety and image quality (Abdullah Suhaimi et al., 2015) although this did not compare DBT to FFDM (or any other screening test). However, the study design of a number of papers may also provide insights into the acceptability of DBT to women. These papers are also discussed in this section.

We found no other articles that specifically investigated the acceptability of DBT to women across the other dimensions of interest to our review. That is, no specific articles or studies investigated acceptability to population sub-groups (including women with a higher than average lifetime risk of breast cancer, younger women or women with more dense breasts). Acceptability to women is an outcome of the PROSPECTS trial (United Kingdom) and an Italian RCT based in Reggio Emilia. Data from the Maroondah trial is also likely to provide information on women's experiences as DBT is implemented. Together, findings from these studies will provide further clarity about women's experiences and the likely acceptability of this screening test.

5.1. What do we know?

As well as the article by Abdullah Suhaimi et al. (2015), a small number of other articles provided commentary about studies in which women had a choice about whether to receive DBT, FFDM alone or FFDM + DBT. Others provided anecdotal commentary about the potential impact of a reduction in recall rates on women's mental health. These articles provide initial inferences about potential acceptability in relation to:

- women's overall choice of screening test
- reduced anxiety due to reduced recall rates for FFDM + DBT compared to FFDM alone, and
- pain related to compression.

5.1.1. Overall choice

The value of decreasing screening recall rates is very high. A decrease in recall rates (especially false positive recall rates) can be directly translated into decreased health costs and less anxiety for women. These benefits are likely to be greatest in younger women, and those with more dense breasts (Haas et al., 2013). No data has been presented on the overall choice of the woman. However, there is some evidence that women may choose either FFDM + DBT or FFDM when given a choice. For example, 88% of study participants in Rose et al.'s 2013 study consented to have FFDM + DBT (other study participants chose to have FFDM). When enrolling in Freer et al.'s study, women who are more informed about DBT either chose to have DBT because they were aware of its cancer detection benefits or chose not to have it because they

were aware of the increased radiation dose associated with dual image acquisition (Freer et al., 2017). While not robust indicators of acceptability to women, these examples demonstrate that, if well-explained or if women are well-informed, they may choose DBT over FFDM or they may not.

5.1.2. Compression and pain/discomfort

Coop et al. (2016) noted that pain associated with compression was a key reason why women may choose not to participate in a screening program. We therefore looked for data describing women's response to compression and DBT compared to FFDM. The most studied DBT system (Hologic) records DBT images under the same compression as FFDM images. DBT images are followed by the FFDM view (NHS, 2014). A small amount of additional time under compression is required to acquire all images (and the amount of time varies by DBT system). As noted by Sechopoulos (2013), DBT reduces tissue overlap. With improved visibility, it is thought that DBT can be performed with reduced compression.

The literature is not settled on the best balance between reduced compression, image quality and women's preference.

Early studies (cited in Coop et al., 2016) indicate that reducing compression from 4cm to 6cm did not adversely affect image quality. Sechopoulos (2013) reported that women did prefer reduced compression (citing Fornvik et al., 2010), but the three participating radiologists did not as image quality was poorer in this study.

Information from the STORM and Malmö prospective trials provides further insight about compression time and its potential impact on women. In a clinical screening setting, Bernardi et al. (2012) reported that the average acquisition time is slightly longer for FFDM + DBT compared to DBT alone: 4m 3s (range = 3m 53s to 4m 18s) compared to 3m 13s (range = 3m 0 s to 3m 26s; p=.01). This means that a woman's breast is under compression for up to one minute longer, which may increase overall pain/discomfort associated with this test. The authors do not report women's feedback on this nor do they report on reader feedback about image quality.

In the Malmö trial, Lång et al. (2016A) performed the DBT with reduced compression compared to FFDM to determine if reduced compression would compromise acceptable image acquisition and cancer detection. They reported that reduced compression of up to 50% was achieved in 90% of cases. Women with larger breasts required more pressure to acquire an acceptable DBT_{MLO} image. Lång et al. reported women's positive feedback about the reduced compression but did not collect specific data on this outcome.

Abdullah Suhaimi et al. (2015) reported on their study of 130 Malaysian women's anxiety during participation in a FFDM + DBT screening examination using Hologic's Selenia Dimensions system. Using a validated questionnaire (State-Trait Anxiety Inventory' Form Y-1), two study radiologists reported a reduction in women's pain and anxiety with reduced compression (38.5 newtons compared to 93.0 newtons for standard compression). They found that the mean anxiety score decreased with reduced compression (from 57.15 to 47.23; p<.001). The mean pain during procedure score reduced from 2.13 to 0.69 (p<.001). The authors noted that image quality (as reported by the two participating radiologists) was not compromised, but no additional data on screening outcomes is provided.

We also reviewed the United Kingdom National Health Service's practical evaluations of two key DBT-capable systems for information on acceptability to women: GE's SenoClaire DBT system and Hologic's Selenia Dimensions DBT system (Bonsall et al., 2016; Mungutroy et al., 2014). We selected these two systems because they reflect the systems used in the evidence base presented

in this report. A practical evaluation for Siemens Mammomat Inspirations system was not available. The practical evaluations for Hologic and GE's DBT systems reported that most radiologists considered compression times to be acceptable (although for the Hologic system 4/10 participating radiologists rate DBT compression time to be "worse" than FFDM alone with the remaining six noting that it was the same as FFDM: the dimensions of "worse" are not explained). Radiologist reports for both the Hologic Dimensions and GE SenoClaire systems were that women's comfort was average to excellent but, for GE SenoClaire, radiologists reported that they had received no feedback from women to indicate that the system was more uncomfortable that FFDM alone (Bonsall et al., 2016; Mungutroy et al., 2014). None of the assessments of women's comfort appear to be validated by women themselves.

No other studies have investigated whether compression and pain/discomfort are issues for women and, if so, whether it would affect their participation in a screening program. However, it appears intuitive that women would appreciate lower compression (and presumably less discomfort) provided the ability to visualise and detect very small cancers was not adversely affected.

5.1.3. Reduced anxiety following participation in a screening program

There is a growing body of evidence to suggest a statistically significant association between FFDM + DBT and a decrease in false recall rates compared to FFDM (see *section 3.2.2*). That is, DBT is a more specific test and enables more accurate interpretation of images with fewer instances of cancer being suspected when none are present. It is logical to assume that a decrease in false recall rate is likely to result in a fall in anxiety experienced by women if they receive a false positive mammogram result.

6. POLICY OR POSITION STATEMENTS ON DBT FROM OTHER JURISDICTIONS

While evidence describing the sensitivity, specificity, safety, longer term mortality benefit, financial cost-effectiveness and acceptability of DBT in a screening setting continues to evolve, articles reviewed for this literature review show that DBT is already being implemented into clinical screening practice in some jurisdictions. This includes community-based radiology practices/single institutions in Taiwan and the USA. Other jurisdictions have developed or recently updated regional or national policy statements or position papers describing their current approach to the use of DBT for screening purposes.

To supplement the scientific evidence base, we completed a grey literature search for national-level position statements. In total, we identified 10 articles and two opinion pieces covering the following jurisdictions:

- Brazil
- Europe (from the European Society of Breast Imaging)
- France
- Italy
- Japan
- New Zealand
- United Kingdom, and
- USA.

An International Agency for Research on Cancer (IARC) working group has also issued a statement on the use of DBT. Statements (and any conclusions from literature updates underpinning the positions) are discussed in *section 6.1*. Key themes are described in *section 6.2*.

6.1. Description of different jurisdictions' advice

6.1.1. Brazil

Brazil is the only jurisdiction studied that specifically recommends DBT for breast cancer screening. The Brazilian College of Radiology and Diagnostic Imaging, the Brazilian Breast Disease Society and the Brazilian Federation of Gynaecological and Obstetrical Associations published a joint set of recommendations for breast cancer screening in Brazil in 2017 (Urban et al., 2017). Based on the OTS and STORM trial and other corroborating results, the authors agreed that the efficacy of DBT in screening for breast cancer has been confirmed because it increases cancer detection rates and decreases overall and false positive recall rates. They noted that the FDA still recommends that DBT be used in combination with FFDM. Concerns around radiation dose were noted but the authors cited evidence that s2DM maintains the benefits of DBT while reducing the radiation dose by nearly half. DBT is recommended either as an adjunct test to FFDM or alone in combination with s2DM for women with an average or high risk of breast cancer. This statement will be reviewed every three years. Recommendations related to DBT were classified into Category B- *"Recommendation based on reasonable scientific evidence, with a consistent consensus among the CBR, SBM, and Febrasgo that this recommendation should be strongly supported*."

6.1.2. European Society of Breast Imaging

The EUSOBI and 30 national breast radiology bodies from around Europe published an online position paper on screening for breast cancer generally in 2016 (Sardanelli et al., 2017). This position paper also discussed the potential of DBT in screening programs. The authors agreed that evidence supporting the use of DBT as a screening tool from prospective trials like the OTS, STORM, STORM-2 and Malmö trials shows better performance compared to FFDM alone, especially for FFDM + DBT (i.e., adjunct screening). For key clinical outcomes, they noted that DBT increases the CDR from 0.5 to 2.7 per 1,000 screening examinations and reduces the recall rate (citing Houssami's 2015 literature review on data and implications of DBT's role in population screening). The position paper also noted the promise of DBT + s2DM as a solution to increased radiation exposure when DBT is performed in combination with FFDM.

The societies concluded that while DBT would likely be the future of "routine mammography" in a screening setting, further statistically significant and clinically relevant evidence of a reduction in the interval cancer rate conferred by DBT is needed before its implementation into national screening programs. The societies also noted that there are initial results on a reduction in interval cancers by McDonald et al. (2016) (as discussed in *section 3.1.3* of this literature review); however, further robust evidence is required. The position paper also noted that the likely increase in reading time would need to be considered before routine implementation. Additional evidence will avoid any potential increase in overdiagnosis and/or costs.

6.1.3. France

In 2016, France undertook a large review of its national breast cancer screening program. A related article (Mayor., 2016) noted that, in France, research showed that the reduction in breast cancer was due more to improved treatment rather than early detection via an organised population-based screening program. France is now moving towards individualised screening based on personal risk and the provision of more detailed information to women by their doctors to support informed consent and choice (as reported by Nelson., 2017). No national statement is available in English and it is not clear what role DBT will have in a personalised screening environment (although in their 2017 review, Liberatore et al. noted that DBT is part of the screening environment in France and Monaco).

6.1.4. IARC Working Group

Experts from 16 countries[17] met at the IARC in November 2014 to discuss different breast cancer screening methods (Lauby-Secretan et al., 2015). The IARC Working Group assessed FFDM + DBT compared with FFDM alone in terms of breast cancer mortality, detection rate, interval cancer rate and proportion of false positive screening outcomes. Based largely on results from the STORM trial, it found that FFDM + DBT increased rates of detection for both in-situ and invasive cancers, and that it may reduce false positive screening outcomes compared to FFDM alone. Evidence of an association with reduced breast cancer mortality was inadequate and the authors noted that the radiation dose received with dual acquisition is increased. The IARC Working Group concluded that more evidence was needed before DBT was considered as a screening tool.

[17] Contributing authors to the Working Group represented Australia, Canada, Chile, Finland, France, Italy, Italy, The Netherlands, New Zealand, Nicaragua, and the USA.

6.1.5. Italy

The Italian College of Breast Radiologists (ICBR), Italian Society of Medical Radiology (SIRM) and Italian Group for Mammography Screening (GISMa) released a joint set of recommendations on DBT in May 2017 (Bernardi et al., 2017). Based on results from the STORM, OTS and Malmö prospective trials and five retrospective studies, the three organisations agreed that FFDM + DBT shows increased cancer detection rates and decreased false positive recall rates compared to FFDM alone. Radiation exposure is still an issue to take into consideration for a generalised adoption of FFDM + DBT for mass population screening and the Group stated that the solution to this is s2DM.

The three organisations stated that in the context of population-based screening programs, a simple increase in sensitivity and overall diagnostic performance of a new tool is not enough on its own for generalised adoption. Evidence from RCTs is needed before introducing new screening tools. The Group suggested that caution be taken around the implementation of DBT in screening programs due to the possibility that a substantial part of the additional cancers detected by DBT could be over-diagnosed lesions, which could result in an increase in over-treatment of abnormalities that may never present with clinical significance.

The Group recommended that generalised adoption of DBT as a primary screening test wait for specific evidence – particularly statistically significant and clinically relevant reduction in interval cancer rates – but that DBT is a promising intervention.

6.1.6. Japan

While not a national position statement, Uematsu (2017) published an article that discussed possible supplemental breast cancer screening for younger women with dense breasts within population-based screening in Japan. Uematsu noted that the sensitivity of mammography decreases with breast density, and that young Japanese women tend to have much higher breast density than older women. Younger Japanese women (40-49 years) were considered unsuitable for mammography because the sensitivity tends to be lower for smaller breast volume and denser breasts. In Japan, the age-specific breast cancer incidence for women in their 40s is the highest. Therefore, Uematsu stated that supplemental screening strategies for this population need to be considered. The results of trials are promising for the implementation of supplemental DBT breast screening because there is evidence that it improves cancer detection and decreases the recall rate. The author noted that while radiation dose is a concern, 2D images can be synthesised from DBT images, which eliminates the double-dose exposure. The author stated that many large-scale institutions and hospitals in Japan have been converting FFDM into DBT over the last few years, and that DBT is a good option for supplemental breast screening in the Japanese screening program. DBT may replace FFDM in a *"well-resourced Japan*", but the author noted that there are still problems to solve before that can occur.

6.1.7. New Zealand

The National Health Committee (NHC) conducted a literature review to prepare an overview of screening in New Zealand (NHC, 2015). DBT was mentioned under the "Emerging Technologies" section as a technology in the early stages of testing and clinical use which may be able to improve diagnostic accuracy and the early detection of breast cancer. The NHC stated that further evidence is required, such as the completion of trials like Malmö before widespread implementation of DBT in routine screening practice should be considered. Since the publication of the NHC Overview, preliminary results have been released from the Malmö trial. It is unclear whether the NHC is planning to update its position on DBT soon.

6.1.8. United Kingdom

The UK NHS Breast Screening Programme (NHSBSP) updated its current position on DBT in March 2016 (Borrelli and Oduko, 2016). Most of the statement relates to the use of DBT in assessment rather than in a screening setting. Instead, it describes what is required before DBT can be used and outlines some of the current uncertainties about DBT's role in the assessment of breast cancer. Before undertaking clinical use, radiologists in the UK are required to attend NHSBSP-recognised training courses, which have a number of required inclusions such as image acquisition and interpretation. Suppliers of DBT systems must also provide specific training and applications to radiographers and radiologists. Borelli and Oduko mainly focused on assessment, and briefly stated that anyone wishing to undertake trials of DBT in relation to screening would need to apply through appropriate channels (for example, the NHS Research and Development Committee) for approval; however, no information on whether DBT should be included in the NHSBSP is provided.

6.1.9. United States

In 2016, the US Preventive Services Task Force (USPSTF) released an update of its 2009 recommendations on screening for breast cancer (Siu et al., 2016). The recommendations apply to asymptomatic women aged 40 years or older with no pre-existing breast cancer or previously diagnosed high-risk breast lesion and women who do not have a high risk for breast cancer. The USPSTF reported that DBT is most often performed in conjunction with conventional FFDM (that is, FFDM + DBT). It expressed concern that this results in doubling the radiation dose to the woman. Although the FDA had approved s2DM at the time of this report, Siu et al. found that study data on the performance of DBT + s2DM was limited to one mammography reading study comparing sensitivity and specificity (Rose et al., 2013) and one prospective clinical trial (Skaane et al., 2014). Siu et al. also reported that there was limited evidence suggesting a slight increase in the risk of breast biopsy for DBT compared with FFDM. The USPSTF concluded that the current evidence was insufficient to assess the benefits and harms of DBT as a primary screening method for breast cancer, and that further research should be conducted before it is considered as a primary screening tool.

6.1.10. Conclusions

Seven countries and one region (Europe excluding the United Kingdom) have current position papers describing views on the role of DBT (either as a stand-alone or adjunct screening tool) in breast cancer screening. Additionally, the IARC has released a position statement which authors from 16 countries contributed to. Except for Brazil, all the position statements report the same conclusion: existing evidence favours FFDM + DBT compared to FFDM alone for key screening outcomes like CDR and recall rates. It is a promising technology that will have some role in the future of screening programs; however, concern remains around the increased radiation dose associated with dual acquisition, the lack of evidence on long-term performance clinical outcomes like interval cancer rates and the impact on longer-term cancer mortality reduction. Currently, all jurisdictions (except Brazil) recommend that further evidence from prospective trials and RCTs be acquired and used to inform decisions about integration into national screening programs. We note a high correlation between the findings of these statements and the evidence base presented in this literature review.

REFERENCES

Abdullah Suhaimi S.A., Mohamed A., Ahmad M., and Chelliah K.K. 'Effects of Reduced Compression in Digital Breast Tomosynthesis on Pain, Anxiety, and Image Quality'. *Malaysian Journal of Medical Sciences* 22, no. 6 (2015): 40–46.

Aujero M.P., Gavenonis S.C., Benjamin R., Zhang Z., and Holt J.S. 'Clinical Performance of Synthesized Two-Dimensional Mammography Combined with Tomosynthesis in a Large Screening Population'. *Radiology* 283, no. 1 (2017): 70–76. https://doi.org/10.1148/radiol.2017162674.

Bernardi D., Belli P., Benelli E., Brancato B., Bucchi L., Calabrese M., Carbonaro L.A., et al. 'Digital Breast Tomosynthesis (DBT): Recommendations from the Italian College of Breast Radiologists (ICBR) by the Italian Society of Medical Radiology (SIRM) and the Italian Group for Mammography Screening (GISMa)'. *La Radiologia Medica* 122, no. 10 (2017A): 723–30. https://doi.org/10.1007/s11547-017-0769-z.

Bernardi D., Caumo F., Macaskill P., Ciatto S., Pellegrini M., Brunelli S., Tuttobene P., et al. 'Effect of Integrating 3D-Mammography (Digital Breast Tomosynthesis) with 2D-Mammography on Radiologists' True-Positive and False-Positive Detection in a Population Breast Screening Trial'. *European Journal of Cancer* 50, no. 7 (2014): 1232–38. https://doi.org/10.1016/j.ejca.2014.02.004.

Bernardi, D, S Ciatto, M Pellegrini, V Anesi, S Burlon, E Cauli, M Depaoli, et al. 'Application of Breast Tomosynthesis in Screening: Incremental Effect on Mammography Acquisition and Reading Time.' *The British Journal of Radiology* 85, no. 1020 (2012): e1174-8. https://doi.org/10.1259/bjr/19385909.

Bernardi, D., and N. Houssami. 'Breast Cancers Detected in Only One of Two Arms of a Tomosynthesis (3D-Mammography) Population Screening Trial (STORM-2)'. *Breast* 32 (2017B): 98–101. https://doi.org/10.1016/j.breast.2017.01.005.

Bernardi D., Macaskill P., Pellegrini M., Valentini M., Fanto C., Ostillio L., Tuttobene P., Luparia A., and Houssami N. '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'. *The Lancet Oncology* 17, no. 8 (2016): 1105–13. https://doi.org/10.1016/S1470-2045%2816%2930101-2.

Bonafede, M.M., V.B. Kalra, J.D. Miller, and L.L. Fajardo. 'Value Analysis of Digital Breast Tomosynthesis for Breast Cancer Screening in a Commercially-Insured US Population'. *ClinicoEconomics and Outcomes Research* 7 (2015): 53–63. https://doi.org/10.2147/CEOR.S76167.

Bonsall J, Akran T, Tsang R, Jones V, Turnbull A, Cornfored E. 2016. *Practical evaluation of GE SenoClaire digital breast tomosynthesis system*. NHS.

Borelli, C and Oduko J. *NHS Breast Screening Programme: current position on use of tomosynthesis* 28 March 2016.

BreastScreen Australia Accreditation Review Committee. 2015. *BreastScreen Australia National Accreditation Standards.*

Carbonaro, L.A., G. Di Leo, P. Clauser, R.M. Trimboli, N. Verardi, M.P. Fedeli, R. Girometti, et al. 'Impact on the Recall Rate of Digital Breast Tomosynthesis as an Adjunct to Digital Mammography in the Screening Setting. A Double Reading Experience and Review of the

Literature'. *European Journal of Radiology* 85, no. 4 (2016): 808–14. https://doi.org/10.1016/j.ejrad.2016.01.004.

Caumo, F., D. Bernardi, S. Ciatto, P. Macaskill, M. Pellegrini, S. Brunelli, P. Tuttobene, et al. 'Incremental Effect from Integrating 3D-Mammography (Tomosynthesis) with 2D-Mammography: Increased Breast Cancer Detection Evident for Screening Centres in a Population-Based Trial'. *Breast* 23, no. 1 (2013): 76–80. https://doi.org/10.1016/j.breast.2013.11.006.

Ciatto S., Houssami N., Bernardi D., Caumo F., Pellegrini M., Brunelli S., Tuttobene P., et al. 'Integration of 3D Digital Mammography with Tomosynthesis for Population Breast-Cancer Screening (STORM): A Prospective Comparison Study'. *The Lancet Oncology* 14, no. 7 (2013): 583–89. https://doi.org/10.1016/S1470-2045%2813%2970134-7.

Conant E.F., Beaber E.F., Sprague B.L., Herschorn S.D., Weaver D.L., Onega T., Tosteson A.N.A., et al. 'Breast Cancer Screening Using Tomosynthesis in Combination with Digital Mammography Compared to Digital Mammography Alone: A Cohort Study within the PROSPR Consortium'. *Breast Cancer Research and Treatment* 156, no. 1 (2016): 109–16. https://doi.org/10.1007/s10549-016-3695-1.

Coop, P, Cowling A, Lawson B. 'Tomosynthesis as a Screening Tool for Breast Cancer: A Systematic Review'. *Radiography* 22, no. 3 (1 August 2016): e190–95. https://doi.org/10.1016/j.radi.2016.03.002.

Dang, Pragya A., Phoebe E. Freer, Kathryn L. Humphrey, Elkan F. Halpern, and Elizabeth A. Rafferty. 'Addition of Tomosynthesis to Conventional Digital Mammography: Effect on Image Interpretation Time of Screening Examinations'. *Radiology* 270, no. 1 (1 January 2014): 49–56. https://doi.org/10.1148/radiol.13130765.

Department of Health and Ageing. 2009. Horizon Scanning Technology. Prioritising Summary: Breast tomosynthesis – a breast cancer screening tool. Update 2009, Commonwealth of Australia, 2009

Destounis S., Arieno A., and Morgan R. 'Initial Experience with Combination Digital Brea St Tomosynthesis plus Full Field Digital Mammography or Full Field Digital Mammography Alone in the Screening Environment'. *Journal of Clinical Imaging Science* 4, no. 1 (2014): 127838. https://doi.org/10.4103/2156-7514.127838.

Durand, Melissa A., Brian M. Haas, Xiaopan Yao, Jaime L. Geisel, Madhavi Raghu, Regina J. Hooley, Laura J. Horvath, and Liane E. Philpotts. 'Early Clinical Experience with Digital Breast Tomosynthesis for Screening Mammography'. *Radiology* 274, no. 1 (2015): 85–92. https://doi.org/10.1148/radiol.14131319.

Freer P.E., Riegert J., Eisenmenger L., Ose D., Winkler N., Stein M.A., Stoddard G.J., and Hess R. 'Clinical Implementation of Synthesized Mammography with Digital Breast Tomosynthesis in a Routine Clinical Practice'. *Breast Cancer Research and Treatment* 166, no. 2 (2017): 501–9. https://doi.org/10.1007/s10549-017-4431-1.

Friedewald S.M., Rafferty E.A., Rose S.L., Durand M.A., Plecha D.M., Greenberg J.S., Hayes M.K., et al. 'Breast Cancer Screening Using Tomosynthesis in Combination with Digital Mammography'. *JAMA - Journal of the American Medical Association* 311, no. 24 (2014): 2499–2507. https://doi.org/10.1001/jama.2014.6095.

Gilbert, Fiona J, Lorraine Tucker, and Ken C Young. 'Digital Breast Tomosynthesis (DBT): A Review of the Evidence for Use as a Screening Tool.' *Clinical Radiology* 71, no. 2 (2016): 141–50. https://doi.org/10.1016/j.crad.2015.11.008.

Greenberg J.S., Javitt M.C., Katzen J., Michael S., and Holland A.E. 'Clinical Performance Metrics of 3D Digital Breast Tomosynthesis Compared with 2D Digital Mammography for Breast Cancer Screening in Community Practice'. *American Journal of Roentgenology* 203, no. 3 (2014): 687–93. https://doi.org/10.2214/AJR.14.12642.

Gur, D., M.L. Zuley, M.I. Anello, G.Y. Rathfon, D.M. Chough, M.A. Ganott, C.M. Hakim, L. Wallace, A. Lu, and A.I. Bandos. 'Dose Reduction in Digital Breast Tomosynthesis (DBT) Screening Using Synthetically Reconstructed Projection Images. An Observer Performance Study.' *Academic Radiology* 19, no. 2 (2012): 166–71. https://doi.org/10.1016/j.acra.2011.10.003.

Haas B.M., Kalra V., Geisel J., Raghu M., Durand M., and Philpotts L.E. 'Comparison of Tomosynthesis plus Digital Mammography and Digital Mammography Alone for Breast Cancer Screening'. *Radiology* 269, no. 3 (2013): 694–700. https://doi.org/10.1148/radiol.13130307.

Hardesty, Lara A. 'Issues to Consider Before Implementing Digital Breast Tomosynthesis Into a Breast Imaging Practice'. *American Journal of Roentgenology* 204, no. 3 (25 February 2015): 681–84. https://doi.org/10.2214/AJR.14.13094.

Hodgson, Robert, Sylvia H Heywang-Kobrunner, Susan C Harvey, Mary Edwards, Javed Shaikh, Mick Arber, and Julie Glanville. 'Systematic Review of 3D Mammography for Breast Cancer Screening.' *Breast (Edinburgh, Scotland)* 27, no. 9213011 (2016): 52–61. https://doi.org/10.1016/j.breast.2016.01.002.

Houssami N. 'Evidence on Synthesized Two-Dimensional Mammography Versus Digital Mammography When Using Tomosynthesis (Three-Dimensional Mammography) for Population Breast Cancer Screening'. *Clinical Breast Cancer*, no. (Houssami) Sydney School of Public Health, Sydney Medical School, University of Sydney, Sydney, Australia (2017). https://doi.org/10.1016/j.clbc.2017.09.012.

Houssami, Nehmat. 'Digital Breast Tomosynthesis (3D-Mammography) Screening: Data and Implications for Population Screening.' *Expert Review of Medical Devices* 12, no. 4 (2015): 377–79. https://doi.org/10.1586/17434440.2015.1028362.

Houssami N., Bernardi D., Pellegrini M., Valentini M., Fanto C., Ostillio L., Tuttobene P., Luparia A., and Macaskill P. '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, no. (Houssami, Macaskill) Sydney School of Public Health (A27), Sydney Medical School, University of Sydney, Sydney 2006, Australia (2017): 94–99. https://doi.org/10.1016/j.canep.2017.01.008.

Houssami N., Macaskill P., Bernardi D., Caumo F., Pellegrini M., Brunelli S., Tuttobene P., et al. 'Breast Screening Using 2D-Mammography or Integrating Digital Breast Tomosynthesis (3D-Mammography) for Single-Reading or Double-Reading - Evidence to Guide Future Screening Strategies'. *European Journal of Cancer* 50, no. 10 (2014): 1799–1807. https://doi.org/10.1016/j.ejca.2014.03.017.

Houssami, Nehmat, Kristina Lang, Daniela Bernardi, Alberto Tagliafico, Sophia Zackrisson, and Per Skaane. 'Digital Breast Tomosynthesis (3D-Mammography) Screening: A Pictorial Review of Screen-Detected Cancers and False Recalls Attributed to Tomosynthesis in Prospective

Screening Trials.' *Breast (Edinburgh, Scotland)* 26, no. 9213011 (2016): 119–34. https://doi.org/10.1016/j.breast.2016.01.007.

Houssami, Nehmat, and Per Skaane. 'Overview of the Evidence on Digital Breast Tomosynthesis in Breast Cancer Detection.' *Breast (Edinburgh, Scotland)* 22, no. 2 (2013): 101–8. https://doi.org/10.1016/j.breast.2013.01.017.

Houssami, Nehmat, and Robin M Turner. 'Rapid Review: Estimates of Incremental Breast Cancer Detection from Tomosynthesis (3D-Mammography) Screening in Women with Dense Breasts.' *Breast (Edinburgh, Scotland)* 30, no. 9213011 (2016): 141–45. https://doi.org/10.1016/j.breast.2016.09.008.

Hunter S.A., Morris C., Nelson K., Snyder B.J., and Poulton T.B. 'Digital Breast Tomosynthesis: Cost-Effectiveness of Using Private and Medicare Insurance in Community-Based Health Care Facilities'. *American Journal of Roentgenology* 208, no. 5 (2017): 1171–75. https://doi.org/10.2214/AJR.16.16987.

Kalra V.B., Wu X., Haas B.M., Forman H.P., and Philpotts L.E. 'Cost-Effectiveness of Tomosynthesis in Annual Screening Mammography'. *American Journal of Roentgenology* 207, no. 5 (2016): 1152–55. https://doi.org/10.2214/AJR.15.14487.

Lang K., Andersson I., Rosso A., Tingberg A., Timberg P., and Zackrisson S. 'Performance of One-View Breast Tomosynthesis as a Stand-Alone Breast Cancer Screening Modality: Results from the Malmo Breast Tomosynthesis Screening Trial, a Population-Based Study'. *European Radiology* 26, no. 1 (2016A): 184–90. https://doi.org/10.1007/s00330-015-3803-3.

Lang K., Nergarden M., Andersson I., Rosso A., and Zackrisson S. 'False Positives in Breast Cancer Screening with One-View Breast Tomosynthesis: An Analysis of Findings Leading to Recall, Work-up and Biopsy Rates in the Malmo Breast Tomosynthesis Screening Trial'. *European Radiology* 26, no. 11 (2016B): 3899–3907. https://doi.org/10.1007/s00330-016-4265-y.

Lauby-Secretan, Béatrice, Chiara Scoccianti, Dana Loomis, Lamia Benbrahim-Tallaa, Véronique Bouvard, Franca Bianchini, and Kurt Straif. 'Breast-Cancer Screening — Viewpoint of the IARC Working Group'. *New England Journal of Medicine* 372, no. 24 (3 June 2015): 2353–58. https://doi.org/10.1056/NEJMsr1504363.

Lee C.I., Cevik M., Alagoz O., Sprague B.L., Tosteson A.N.A., Miglioretti D.L., Kerlikowske K., et al. 'Comparative Effectiveness of Combined Digital Mammography and Tomosynthesis Screening for Women with Dense Breasts'. *Radiology* 274, no. 3 (2014): 772–80. https://doi.org/10.1148/radiol.14141237.

Liberatore, Mathieu, Jean-Michel Cucchi, Martine Fighiera, Anne Binet, Marie Christine Missana, Philippe Brunner, Michel Yves Mourou, and Antoine Iannessi. 'Interest of Systematic Tomosynthesis (3D Mammography) with Synthetic 2D Mammography in Breast Cancer Screening.' *Hormone Molecular Biology and Clinical Investigation* 32, no. 2 (16 December 2017). https://doi.org/10.1515/hmbci-2017-0024.

Lourenco A.P., Barry-Brooks M., Baird G.L., Tuttle A., and Mainiero M.B. 'Changes in Recall Type and Patient Treatment Following Implementation of Screening Digital Breast Tomosynthesis'. *Radiology* 274, no. 2 (2015): 337–42.

Maxwell, A.J., M. Michell, Y.Y. Lim, S.M. Astley, M. Wilson, E. Hurley, D.G. Evans, et al. 'A Randomised Trial of Screening with Digital Breast Tomosynthesis plus Conventional Digital 2D Mammography versus 2D Mammography Alone in Younger Higher Risk Women'. *European Journal of Radiology* 94 (2017): 133–39. https://doi.org/10.1016/j.ejrad.2017.06.018.

Mayor S. 2016. 'Reduction in breast cancer deaths is due to treatment not screening' in BMJ 2016; 355; doi: https://doi.org/10.1136/bmj.i5544.

McCarthy A.M., Kontos D., Synnestvedt M., Tan K.S., Heitjan D.F., Schnall M., and Conant E.F. 'Screening Outcomes Following Implementation of Digital Breast Tomosynthesis in a General-Population Screening Program'. *Journal of the National Cancer Institute* 106, no. 11 (2014): dju316. https://doi.org/10.1093/jnci/dju316.

McDonald, Elizabeth S, Anne Marie McCarthy, Amana L Akhtar, Marie B Synnestvedt, Mitchell Schnall, and Emily F Conant. 'Baseline Screening Mammography: Performance of Full-Field Digital Mammography Versus Digital Breast Tomosynthesis.' *AJR. American Journal of Roentgenology* 205, no. 5 (2015): 1143–48. https://doi.org/10.2214/AJR.15.14406.

McDonald E.S., Oustimov A., Weinstein S.P., Synnestvedt M.B., Schnall M., and Conant E.F. 'Effectiveness of Digital Breast Tomosynthesis Compared With Digital Mammography: Outcomes Analysis From 3 Years of Breast Cancer Screening'. *JAMA Oncology* 2, no. 6 (2016): 737–43. https://doi.org/10.1001/jamaoncol.2015.5536.

Miller J.D., Bonafede M.M., Herschorn S.D., Pohlman S.K., Troeger K.A., and Fajardo L.L. 'Value Analysis of Digital Breast Tomosynthesis for Breast Cancer Screening in a US Medicaid Population'. *Journal of the American College of Radiology* 14, no. 4 (2017): 467. https://doi.org/10.1016/j.jacr.2016.11.019.

Mungutroy EHL, Oduko JM, Cooke JC , Formstone, WJ. 2014. Practical evaluation of Hologic Selenia Dimensions digital breast tomosynthesis system. NHS.

Nelson R. 2017. Make screening mammography personal say the French. https://www.medscape.com/viewarticle/888005#vp_1. Accessed 18 February 2018.

NHC. (2015). Position statement pamphlet on breast density and screening.

Pan H.-B., Wong K.-F., Yao A., Hsu G.-C., Chou C.-P., Liang H.-L., Huang J.-S., Li H.-J., Wang S.-C., and Yang T.-L. 'Breast Cancer Screening with Digital Breast Tomosynthesis - 4 Year Experience and Comparison with National Data'. *Journal of the Chinese Medical Association*, no. (Pan) Kaohsiung Veterans General Hospital Tainan Branch, Tainan, Taiwan, ROC (2017). https://doi.org/10.1016/j.jcma.2017.05.013.

Poplack, Steven. 'Breast Tomosynthesis: Clinical Evidence.' *Radiologic Clinics of North America* 55, no. 3 (2017): 475–92. https://doi.org/10.1016/j.rcl.2016.12.010.

Powell J.L., Hawley J.R., Lipari A.M., Yildiz V.O., Erdal B.S., and Carkaci S. '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'. *Academic Radiology* 24, no. 3 (2017): 302–7. https://doi.org/10.1016/j.acra.2016.10.001.

Olgar T., Kahn T., and Gosch D. 'Average Glandular Dose in Digital Mammography and Breast Tomosynthesis.' *Fortschr Rontgenstr* 184, no. 10 (2012): 911-918. https://doi.org/10.1055/s-0032-1312877.

Rafferty E.A., Rose S.L., Miller D.P., Durand M.A., Conant E.F., Copit D.S., Friedewald S.M., et al. 'Effect of Age on Breast Cancer Screening Using Tomosynthesis in Combination with Digital Mammography'. *Breast Cancer Research and Treatment* 164, no. 3 (2017): 659–66. https://doi.org/10.1007/s10549-017-4299-0.

Rafferty, Elizabeth A., Jeong Mi Park, Liane E. Philpotts, Steven P. Poplack, Jules H. Sumkin, Elkan F. Halpern, and Loren T. Niklason. 'Assessing Radiologist Performance Using Combined Digital

Mammography and Breast Tomosynthesis Compared with Digital Mammography Alone: Results of a Multicenter, Multireader Trial'. *Radiology* 266, no. 1 (1 January 2013): 104–13. https://doi.org/10.1148/radiol.12120674.

Rodriguez-Ruiz A., Gubern-Merida A., Imhof-Tas M., Lardenoije S., Wanders A.J.T., Andersson I., Zackrisson S., et al. 'One-View Digital Breast Tomosynthesis as a Stand-Alone Modality for Breast Cancer Detection: Do We Need More?' *European Radiology*, no. (Rodriguez-Ruiz, Gubern-Merida, Imhof-Tas, Lardenoije, Karssemeijer, Mann, Sechopoulos) Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, Geert Grooteplein 10, Post 766, Nijmegen 6525 GA, Netherlands (2017): 1–11. https://doi.org/10.1007/s00330-017-5167-3.

Rose S.L., Tidwell A.L., Ice M.F., Nordmann A.S., Sexton R., and Song R. 'A Reader Study Comparing Prospective Tomosynthesis Interpretations with Retrospective Readings of the Corresponding FFDM Examinations'. *Academic Radiology* 21, no. 9 (2014): 1204–10. https://doi.org/10.1016/j.acra.2014.04.008.

Rose S.L., Tidwell A.L., Bujnoch L.J., Kushwaha A.C., Nordmann A.S., and Sexton Jr. R. 'Implementation of Breast Tomosynthesis in a Routine Screening Practice: An Observational Study'. *American Journal of Roentgenology* 200, no. 6 (2013): 1401–8. https://doi.org/10.2214/AJR.12.9672.

Rosso A., Lang K., Petersson I.F., and Zackrisson S. 'Factors Affecting Recall Rate and False Positive Fraction in Breast Cancer Screening with Breast Tomosynthesis - A Statistical Approach'. *Breast* 24, no. 5 (2015): 680–86. https://doi.org/10.1016/j.breast.2015.08.007.

Sardanelli, F., H.S. Aase, M. Álvarez, E. Azavedo, H.J. Baarslag, C. Balleyguier, P.A. Baltzer, et al. 'Position Paper on Screening for Breast Cancer by the European Society of Breast Imaging (EUSOBI) and 30 National Breast Radiology Bodies from Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Israel, Lithuania, Moldova, The Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Spain, Sweden, Switzerland and Turkey'. *European Radiology* 27, no. 7 (2017): 2737–43. https://doi.org/10.1007/s00330-016-4612-z.

Sechopoulos, Ioannis. 'A Review of Breast Tomosynthesis. Part I. The Image Acquisition Process.' *Medical Physics* 40, no. 1 (2013): 014301. https://doi.org/10.1118/1.4770279.

Sharpe, R.E., S. Venkataraman, J. Phillips, V. Dialani, V.J. Fein-Zachary, S. Prakash, P.J. Slanetz, and T.S. Mehta. '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, no. 3 (2016): 698–706. https://doi.org/10.1148/radiol.2015142036.

Shin, Sung Ui, Jung Min Chang, Min Sun Bae, Su Hyun Lee, Nariya Cho, Mirinae Seo, Won Hwa Kim, and Woo Kyung Moon. 'Comparative Evaluation of Average Glandular Dose and Breast Cancer Detection between Single-View Digital Breast Tomosynthesis (DBT) plus Single-View Digital Mammography (DM) and Two-View DM: Correlation with Breast Thickness and Density'. *European Radiology* 25, no. 1 (1 January 2015): 1–8. https://doi.org/10.1007/s00330-014-3399-z.

Siu, Albert L. 'Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement.' *Annals of Internal Medicine* 164, no. 4 (16 February 2016): 279–96. https://doi.org/10.7326/M15-2886.

Skaane, P. 'Breast Cancer Screening with Digital Breast Tomosynthesis'. *Breast Cancer* 24, no. 1 (2017): 32–41. https://doi.org/10.1007/s12282-016-0699-y.

Skaane P., Bandos A.I., Eben E.B., Jebsen I.N., Krager M., Haakenaasen U., Ekseth U., Izadi M., Hofvind S., and Gullien R. 'Two-View Digital Breast Tomosynthesis Screening with Synthetically Reconstructed Projection Images: Comparison with Digital Breast Tomosynthesis with Full-Field Digital Mammographic Images'. *Radiology* 271, no. 3 (2014): 655–63. https://doi.org/10.1148/radiol.13131391.

Skaane P., Bandos A.I., Gullien R., Eben E.B., Ekseth U., Haakenaasen U., Izadi M., et al. 'Prospective Trial Comparing Full-Field Digital Mammography (FFDM) versus Combined FFDM and Tomosynthesis in a Population-Based Screening Programme Using Independent Double Reading with Arbitration'. *European Radiology* 23, no. 8 (2013A): 2061–71. https://doi.org/10.1007/s00330-013-2820-3.

Skaane, Per, Andriy I Bandos, Randi Gullien, Ellen B Eben, Ulrika Ekseth, Unni Haakenaasen, Mina Izadi, et al. 'Comparison of Digital Mammography Alone and Digital Mammography plus Tomosynthesis in a Population-Based Screening Program.' *Radiology* 267, no. 1 (2013B): 47–56. https://doi.org/10.1148/radiol.12121373.

Starikov A., Drotman M., Hentel K., Katzen J., Min R.J., and Arleo E.K. '2D Mammography, Digital Breast Tomosynthesis, and Ultrasound: Which Should Be Used for the Different Breast Densities in Breast Cancer Screening?' *Clinical Imaging* 40, no. 1 (2016): 68–71. https://doi.org/10.1016/j.clinimag.2015.10.001.

Strudley CJ, Looney P, Young KC. 2014. *Technical evaluation of Hologic Selenia Dimensions digital breast tomosynthesis system*. NHS.

Strudley CJ, Warren P, Young KC. 2015. *Technical evaluation of Siemens Mammomat Inspirations digital breast tomosynthesis system*. NHS.

Sumkin J.H., Ganott M.A., Chough D.M., Catullo V.J., Zuley M.L., Shinde D.D., Hakim C.M., Bandos A.I., and Gur D. 'Recall Rate Reduction with Tomosynthesis During Baseline Screening Examinations: An Assessment From a Prospective Trial'. *Academic Radiology* 22, no. 12 (2015): 1477–82. https://doi.org/10.1016/j.acra.2015.08.015.

Svahn, T M, N Houssami, I Sechopoulos, and S Mattsson. 'Review of Radiation Dose Estimates in Digital Breast Tomosynthesis Relative to Those in Two-View Full-Field Digital Mammography.' *Breast (Edinburgh, Scotland)* 24, no. 2 (2015A): 93–99. https://doi.org/10.1016/j.breast.2014.12.002.

Svahn, Tony M, Petra Macaskill, and Nehmat Houssami. '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 (Edinburgh, Scotland)* 24, no. 6 (2015B): 687–93. https://doi.org/10.1016/j.breast.2015.08.012.

Uematsu T. 'The Need for Supplemental Breast Cancer Screening Modalities: A Perspective of Population-Based Breast Cancer Screening Programs in Japan'. *Breast Cancer* 24, no. 1 (2017): 26–31. https://doi.org/10.1007/s12282-016-0707-2.

Urban, L.A.B.D., L.F. Chala, S.P. Bauab, M.B. Schaefer, R.P. Santos, N.M.A. Maranhão, A.L. Kefalas, et al. 'Breast Cancer Screening: Updated Recommendations of the Brazilian College of Radiology and Diagnostic Imaging, Brazilian Breast Disease Society, and Brazilian Federation of Gynecological and Obstetrical Associations'. *Radiologia Brasileira* 50, no. 4 (2017): 244–49. https://doi.org/10.1590/0100-3984.2017-0069.

Vedantham S., Karellas A., Vijayaraghavan G.R., and Kopans D.B. 'Digital Breast Tomosynthesis: State of the Art1'. *Radiology* 277, no. 3 (2015): 663–84. https://doi.org/10.1148/radiol.2015141303.

Wang W.-S., Hardesty L., Borgstede J., Takahashi J., and Sams S. 'Breast Cancers Found with Digital Breast Tomosynthesis: A Comparison of Pathology and Histologic Grade'. *Breast Journal* 22, no. 6 (2016): 651–56. https://doi.org/10.1111/tbj.12649.

Yaffe, Martin J. 'Reducing Radiation Doses for Breast Tomosynthesis?' *The Lancet Oncology* 17, no. 8 (n.d.): 1027–29. https://doi.org/10.1016/S1470-2045(16)30155-3.

Zuckerman, Samantha P., Emily F. Conant, Brad M. Keller, Andrew D. A. Maidment, Bruno Barufaldi, Susan P. Weinstein, Marie Synnestvedt, and Elizabeth S. McDonald. 'Implementation of Synthesized Two-Dimensional Mammography in a Population-Based Digital Breast Tomosynthesis Screening Program'. *Radiology* 281, no. 3 (28 July 2016): 730–36. https://doi.org/10.1148/radiol.2016160366.

Zuley, Margarita L., Ben Guo, Victor J. Catullo, Denise M. Chough, Amy E. Kelly, Amy H. Lu, Grace Y. Rathfon, et al. 'Comparison of Two-Dimensional Synthesized Mammograms versus Original Digital Mammograms Alone and in Combination with Tomosynthesis Images'. *Radiology* 271, no. 3 (2014): 664–71. https://doi.org/10.1148/radiol.13131530.

Zuley M.L., Bandos A.I., Abrams G.S., Cohen C., Hakim C.M., Sumkin J.H., Drescher J., Rockette H.E., and Gur D. 'Time to Diagnosis and Performance Levels during Repeat Interpretations of Digital Breast Tomosynthesis. Preliminary Observations'. *Academic Radiology* 17, no. 4 (2010): 450–55. https://doi.org/10.1016/j.acra.2009.11.011.

APPENDIX A: COMBINED EVIDENCE TABLES

Combined evidence table for cancer detection rates

Study	Sample	Study type	DBT + s2DM **CDR per 1000 screening examinations (95%CI; p-value)**	FFDM + DBT **CDR per 1000 screening examinations (95%CI; p-value)**	FFDM alone **CDR per 1000 screening examinations (95%CI; p-value)**	DBT + s2DM **Invasive CDR per 1000 screening examinations (95%CI; p=value)**	FFDM + DBT **Invasive CDR per 1000 screening examinations (95%CI; p-value)**	FFDM alone **Invasive CDR per 1000 screening examinations (95%CI; p=value)**	Incremental detection **(95%CI)**	FFDM + DBT / FFDM alone **Invasive CDR per 1000 screening examinations (95%CI; p=value)**	**PPV**
Prospective studies											
Ciatto et al., 2013 (STORM) *Main article*	7292 asymptomatic, average risk Italian women aged 48 years or older (median age 58 years)	Prospective, fully paired trial using Hologic Selenia Dimensions systems in combination mode		8.1 (6·2–10·4) (p<0.0001 compared to FFDM alone)	5.3 (3·8–7·3)		7.1	4.8	2.7 (1.7-4.2) 53% increase (p=.84)		
Caumo et al., 2014 (STORM)		Evaluation of screening metrics at two sites: Trento Verona		7.8 (5.3-10.9) 8.6 (5.6-12.5) (p=.79)	4.9 (3.1-7.5) 5.9 (3.5-9.3) (p=.63)		NR	NR	2.8 (1.5-4.9) 2.6 (1.1-5.2) (p=1)		
Houssami et al., 2014 (STORM)		Analysis of STORM data using different reading strategies: Single reading Double reading		7.5 (5.7-9.8; p=.001) 8.1 (6.2-10.4; p=.001)	4.8 (3.3-6.7) 5.3 (3.8-7.3)		NR	NR	2.7 (1.6-4.2) 2.7 (1.6-4.2)		
Bernardi et al., 2016 (STORM-2)	9672 asymptomatic Italian women aged 49 years or older (median age 58 years) who attended population-based screening	Prospective, fully paired trial using Selenia Dimensions system with C-view for FFDM + DBT with single reading (*NB analysis of 9677 women*)	8.8 (7.0 to 10.8; p<.0001 compared to FFDM; (p=.58 compared to FFDM + DBT)	8.5 (6.7 to 10.5; p<.0001 compared to FFDM)	6.3 (4.8 to 8.1)						

Lång et al., 2016A (Malmö trial)				8.9 (6.9-11.3; *p*<.0001)	6.3 (4.6-8.3; *p*<.0001)						24% for FFDM and DBT
Skaane et al., 2013 (OTS trial)	12,631 Norwegian women aged 50-69 years (average age = 59.3) participating in the biennial Oslo breast screening program, with nine months follow-up.	Prospective, fully paired trial using Hologic Dimensions system with double reading		8.0	6.1		6.4	4.4	40% increase in detection of invasive cancers (*p*<0.001)		
Skaane et al., 2013 (OTS trial)		Prospective, fully paired trial using paired analysis of imaging arms		9.4	7.1		NR	NR	30% increase (*p*<0.001)		PPV: (*p*=.72) 28.5% FFDM alone 29.1% FFDM + DBT
Skaane et al., 2014 (OTS trial)	12,270 screens from 24,901 Norwegian women aged 50-69 years (mean age 59.2 years)	Prospective, fully paired trial using Selenia Dimensions system with C-view for FFDM + DBT with double reading and two study periods (one using C-view, one using an earlier software. Only rates using C-view are reported here)	Study period 1: 7.4 Study period 2: 7.8 Study period 1 = decrease of 7% Study period 2 = decrease of 2%	Study period 1: 8.0 Study period 2: 7.7							Study period 1: PPV_1 (*p*=.61) 30.3% DBT + s2DM 28.5% FFDM + DBT Study period 2: PPV_1 (*p*=.47) 34.9% DBT + s2DM 32.1% FFDM + DBT
Retrospective studies											
Aujero et al., 2017	Mammograms from a single USA practice: 16,173 mammograms with DBT + s2DM; 30,561 mammograms with FFDM + DBT; 32,076 mammograms with FFDM alone	Retrospective observational study with single reading using Selenia Dimensions system with C-view	6.1 (*p*=.27 compared to FFDM; *p*=.71 compared to FFDM + DBT)	6.4 (OR, 1.21; 0.98-1.48)	5.3	76.5% (*p*=<.01 compared to FFDM + DBT)				FFDM + DBT: 61.3%	PPV_1: (*p*=.001) 14.3% DBT + s2DM 10.9% FFDM + DBT 6% FFDM alone PPV_2: (*p*=.01) 39.3% DBT + s2DM 26.3% FFDM + DBT 20.9% FFDM alone PPV_3: (*p*=.001) 40.8% DBT + s2DM

											28.5% FFDM + DBT 22.2% FFDM alone
Freer et al., 2017	31,979 women receiving a screening mammogram a single USA practice between 10/2013–12/2015 (9525 women screened with DBT + s2DM; 1019 screened with FFDM + DBT; 21,435 screened with FFDM alone	Retrospective analysis using Hologic Selenia and Dimensions systems with C-view	5.9 (non-adjusted) 5.4 (adjusted)	6.9 (non-adjusted) 5.7 (adjusted)	5.9 (non-adjusted) 5.0 (adjusted) (For adjusted CDR: *p*=.66 compared to FFDM; *p*=.90 compared to FFDM + DBT	4.6 (non-adjusted) 4.3 (adjusted)				*Non-adjusted* FFDM + DBT: 3.9 FFDM alone: 4.3 *Adjusted* FFDM + DBT: 3.4 FFDM alone: 3.9	*Adjusted PPV* PPV_1: 9.1% DBT + s2DM 8.1% FFDM + DBT 6.2% FFDM alone PPV_2: (*p*=.054) 36.4% FFDM + DBT 40.3% DBT + s2DM 30.9% FFDM alone PPV_3: (*p*=.53) 36.7% FFDM + DBT 36.3% DBT + s2DM 31% FFDM alone
Pan et al., 2017	No specific description of the sample provided	Retrospective analysis comparing screening outcomes before/after implementation of DBT (Hologic Dimensions system) from a single hospital site to national data from Taiwan's National Cancer Registry		2012: 8.5 2013: 10.1 2014: 11.4 2015: 8.7	2009: 6.3 2010: 8.1 2011: 7.5		NR	NR	Average of 32.2% increase		Average PPV_1: 10.1% FFDM + DBT 6.1% FFDM Average PPV_2: 33.27% FFDM + DBT 31.0% FFDM Average PPV_3: 38.47% FFDM + DBT 38.5% FFDM
Powell et al., 2017	FFDM + DBT: 2304 FFDM: 10,477	Retrospective observational data review of images generated with Hologic's Selenia + Dimensions systems		7.8	5.2		3.5 (1.5 to 6.8) (*p*=.805 compared to FFDM)	3.1 (2.2 to 4.4)	12% difference in invasive CDR		
Rafferty et al., 2017 (PROSPR consortium centres)	FFDM + DBT: 173,414 FFDM: 278,906	Retrospective, multicentre analysis of images taken using Hologic's Selenia Dimensions system									PPV_1: 40-49y: (*p*=.001) 3.4% FFDM + DBT 2.3% FFDM alone 50-59y: (*p*=.001) 6.0% FFDM + DBT 3.8% FFDM alone 60-69y: (*p*=.001) 10.3% FFDM + DBT 6.9% FFDM alone

											70+: (*p*=.003) 12.6% FFDM + DBT 9.6% FFDM alone PV_3: 40-49y: (*p*=.006) 17.6% FFDM + DBT 13.7% FFDM alone 50-59y: (*p*=.012) 26.3% FFDM + DBT 21.9% FFDM alone 60-69y: (*p*=.077) 39.2% FFDM + DBT 35% FFDM alone 70+: (*p*=.55) 43.0% FFDM + DBT 45.1% FFDM alone
Conant et al., 2016	FFDM + DBT: 55,998 FFDM: 142,883 Women aged 40 to 74 years	Retrospective analysis of data from three PROSPR consortium sites (NB mammography system used not stated)		6.5 (adjusted OR 1.49; 95%CI=1.17 to 1.89; (*p*=.0016)	4.9		4.7 (adjusted OR 1.45; 95%CI=1.09 to 1.92; *p*=.0252)	3.7	34% increase in all cancers, 27% increase in invasive cancer		PPV_1: (*p*=.0001) 6.4% FFDM + DBT 4.1% FFDM
McDonald et al., 2016	FFDM + DBT: 33,740 FFDM: 10,728 *12079 had one screen* *6293 had two screens* *3023 had three screens*	Retrospective review of mammography metrics from a single site over four years of screening with single reading using Hologic Dimensions system		6.1 (Year 3) 5.8 (Year 2) 5.5 (Year 1) (*p*=.60 compared to FFDM alone)	4.6		NR	NR	34.1% increase		PPV_1 (Year 0/Year3) (*p*=.02): 6.7% FFDM + DBT 4.4% FFDM
Sharpe et al., 2016	FFDM + DBT: 5703 FFDM: 80,149	Prospective study with a retrospective cohort performed at a single site using Hologic Dimensions system for DBT and GE Senographe Essential, 2000D and DS systems		5.4 (3.7 to 7.8)	3.5 (3.1 to 3.9)		2.81 (*p*=.61 compared to FFDM)	2.46	54.3% increase (*p*<.0018)		

Wang et al., 2016	FFDM + DBT: 12,444 FFDM: 12,444	Retrospective study of DBT and FFDM images (Hologic Dimensions system) of 65 breast cancers		5.2	4.4		3.3	2.6	18% increase in all cancers; increase of 27% for invasive		
Zuckerman et al., 2016	FFDM + DBT: 15,571 DBT + s2DM: 5366	Observational study set in a community screening setting using Hologic Dimensions system	5.03 (p=.72 compared to FFDM + DBT)	5.45	NA	3.85				4.10 (p=.84 compared to FFDM + DBT)	PPV_1: (p=.58) 7.1% DBT + s2DM 6.2% FFDM + DBT PPV_2: (p=.054) 35.5% DBT + s2DM 24.7% FFDM + DBT PPV_3: (p=.53) 38.6% DBT + s2DM 27% FFDM + DBT
Durand et al., 2015	FFDM + DBT: 8591 FFDM: 9364	Retrospective review which includes CAD using Hologic Selenia and Dimensions systems		5.9 (p=.88 compared to FFDM alone) (p=.12 compared to historical control)	5.7 4.4 (historical control)		NR	NR	NR		
Lourenco et al., 2015	FFDM + DBT: 12,921 (ages 30.9-89.4 years, average age = 54.6 years) FFDM: 12,577 (ages 29.4-90.6 years, average age = 55.3 years)	Retrospective review of two cohorts (DBT alone=2012/13, FFDM=2011/12), single reading with CAD. FFDM performed using GE Senographe series. DBT performed with Hologic Selenia Dimensions system.		DBT alone 4.6 (p=.44)	5.4 (p=.44)						PPV_3: (p=.21) 30.2% FFDM 23.8 DBT
McDonald et al., 2015	FFDM + DBT (Prevalent): 1859 FFDM + DBT (Incident): 9524	Observational study using Hologic Dimensions system for women presenting for		Prevalent: 5.4 (p=.41) Incident: 5.9	Prevalent: 4.6 Incident: 4.2		NR	NR	Prevalent: 40.5% Incident: 17.4% (p=.74)		PPV_1: prevalent (p=.25) 3.7% FFDM + DBT 2.0% FFDM PPV_1: incident (p=.09)

	FFDM (Prevalent): 1204 FFDM (Incident): 13,712	mammographic screening at a single institution *NB Data comes from McCarthy et al., 2015*		(*p*=.51)							6.9% FFDM + DBT 5.1% FFDM PPV_2: prevalent (*p*=.81) 12.8% FFDM + DBT 14.5% FFDM PPV_2: incident (*p*=.48) 27.6% FFDM + DBT 24.6% FFDM PPV_3: prevalent (*p*=.84) 14.1% FFDM + DBT 15.6% FFDM PPV_3: incident (*p*=.65) 28.7% FFDM + DBT 26.6% FFDM (incident)
Destounis et al., 2014	524 women aged >30 years (mean age 59 years) including women with a history of breast cancer	Retrospective review of images with double reading. FFDM system was Hologic Selenia or Dimensions, GE Senographe Essential or Fuji CRm. DBT system was Hologic Selenia Dimensions.		5.7	3.8						PPV_3: 16.7% FFDM 50.0% FFDM + DBT
Friedewald et al., 2014	FFDM + DBT: 173,663 (ages 52.6-59.7 years, average age = 56.2 years) FFDM: 281,187 (ages 54.4-60.5 years, average age = 57.0 years)	Retrospective review with single reader using data from 13 centres all using Hologic Selenia Dimensions systems		5.4 (4.9 to 6.0; *p*<.001 compared to FFDM alone)	4.2 (3.8 to 4.7)		4.1 (3.7 to 4.5; *p*<.001 compared to FFDM alone)	2.9 (2.5 to 3.2)	28.6% increase		Mean PPV_1: (*p*<.001) 6.1% FFDM + DBT 4.1% FFDM Mean PPV_3: (*p*<.001) 29.2% FFDM + DBT 24.2% FFDM
Greenburg et al., 2014	FFDM + DBT: 20,943 FFDM: 38,674 No differences in study arms by age, ethnicity, family history of BC, or	Retrospective review of mammography outcomes at a multi-site radiology practice using Hologic Selenia or Selenia Dimensions systems		6.3 (*p*=.348)	4.9		4.6 (*p*=.0056)	3.2	28.6% increase (*p*=.035)		PPV_1: (*p*=.0003) 4.6% FFDM + DBT 3.0% FFDM PPV_3: 22.7% FFDM + DBT 21.5% FFDM

	prevalence or incidence screening										
McCarthy et al., 2014	FFDM + DBT: 15,571 (average age = 56.7 years) FFDDM: 10,728 (average age = 56.9 years)	Observational study using Hologic Dimensions system for women presenting for mammographic screening at a single institution *NB McDonald et al., (2015) reported on baseline/prevalence CDR based on this study*		5.5 (4.3 to 6.6)	4.6 (3.3 to 5.8)		3.9 (2.9 to 4.8)	3.2 (2.1 to 4.2)	19.6% increase in total CDR (0.9 per 1000 screening examinations (*p*=.32) 21.9% increase in invasive CDR (*p*=.36)		PPV_1: (*p*=.047) 6.2% FFDM + DBT 4.4% FFDM PPV_2: 24.7% FFDM + DBT 22.4% FFDM PPV_3: 25.4% FFDM + DBT 24.7% FFDM
Haas et al., 2013	FFDM + DBT: 6100 FFDM alone: 7058	Retrospective analysis using Hologic Selenia and Dimensions systems *NB: No rate is statistically significant*		Average risk: 5.7 Increased risk: 8.6 Baseline risk: 5.1	Average risk: 5.2 Increased risk: 7.9 Baseline risk: 4.5		NR	NR	NR		
Rose et al., 2013	FFDM + DBT images: 10,878 (average age = 54.5 years) FFDM images: 10,878 (average age = 53.8 years)	Observational reading study of data before/after DBT implemented using Hologic Selenia and Dimensions systems		5.4 (*p*<.0001)	3.5		4.3 (*p*=.07)	2.8	66% increase (*p*<0.0001)		PPV_1: 10.1% FFDM + DBT 4.7% FFDM Average PPV_3: 39.8% FFDM + DBT 26.5% FFDM

Combined evidence table for recall rates

Study	Sample	Study type	DBT + s2DM **Recall rates (95%CI; *p*-value)**	FFDM + DBT **Recall rate (95%CI; p-value)**	FFDM alone **Recall rate (95%CI; p-value)**	Difference between recall rates **(95%CI; p-value)**	DBT + s2DM **False positive rate (95%CI; p-value)**	FFDM + DBT **FPR (95%CI; p-value)**	FFDM alone **FPR (95%CI; p-value)**	Difference between FPR **(95%CI; p-value)**
Prospective trials										
Ciatto et al., 2013 (STORM) *Main article*	7292 asymptomatic, average risk Italian women aged 48 years or older (median age 58 years)	Prospective, fully paired trial using Hologic Selenia Dimensions systems in combination mode		4.3%	5.0%	NR		3.5%	4.4%	9.3 decrease per 1000 screens (-11.8 to -7.2)
Caumo et al., 2014 (STORM)		Evaluation of screening metrics at two sites: Trento Verona								
Houssami et al., 2014 (STORM)		Analysis of STORM data using different reading strategies: Single reading Double reading								
Bernardi et al., 2016 (STORM-2)	9672 asymptomatic Italian women aged 49 years or older (median age 58 years) who attended population-based screening	Prospective, fully paired trial using Selenia Dimensions system with C-view for FFDM + DBT with single reading (*NB analysis of 9677 women*)					4.45% (statistically significant increase vs. FFDM: *p*<.001; and FFDM + DBT: *p*=.03)	3.97% (statistically significant increase vs. FFDM: *p*<.001)	3.42%	
Lång et al., 2016A (Malmö trial)				3.8% (3.3 to 4.2)	2.6% (2.3 to 3.0)	Increase in recall rate using DBT relative to DM: 43% (26 to 62; *p*<.0001)				

Skaane et al., 2013 (OTS trial)	12,631 Norwegian women aged 50-69 years (average age = 59.3) participating in the biennial Oslo breast screening program, with nine months follow-up.	Prospective, fully paired trial using Hologic Dimensions system with double reading		3.67%	2.9%	NR		8.5%	10.3%	Pre-arbitration: 8 decrease per 1000 screens Post-arbitration: 5.4 increase per 1000 screens (4.2 to 6.8)
Skaane et al., 2013 (OTS trial)		Prospective, fully paired trial using paired analysis of imaging arms		2.78%	2.1%	NR				
Skaane et al., 2014 (OTS trial)	12,270 screens from 24,901 Norwegian women aged 50-69 years (mean age 59.2 years)	Prospective, fully paired trial using Selenia Dimensions system with C-view for FFDM + DBT with double reading and two study periods (one using C-view, one using an earlier software. Only rates using C-view are reported here)					4.5% (non-significant increase vs. FFDM + DBT: p=.85)	4.6%	NR for sub-analysis	
Retrospective trials										
Aujero et al., 2017	Mammograms from a single USA practice: 16,173 mammograms with DBT + s2DM; 30,561 mammograms with FFDM + DBT; 32,076 mammograms with FFDM alone	Retrospective observational study with single reading using Selenia Dimensions system with C-view	4.3% (statistically significant decrease vs. FFDM + DBT and FFDM: p<.001)	5.8% (statistically significant decrease vs. FFDM: p<.001)	8.7%		3.6% (statistically significant decrease vs. FFDM + DBT and FFDM: p<.001)	5.2% (statistically significant decrease vs. FFDM: p<.001)	8.2%	
Freer et al., 2017	31,979 women receiving a screening mammogram a single USA practice between 10/2013–	Retrospective analysis using Hologic Selenia and Dimensions systems with C-view	5.52% (decrease vs. FFDM + DBT non-significant:	6.39% (statistically significant decrease vs.	7.83%					

	12/2015 (9525 women screened with DBT + s2DM; 1019 screened with FFDM + DBT; 21,435 screened with FFDM alone		*p*=.25)	FFDM: *p*<.001)						
Pan et al., 2017	No specific description of the sample provided	Retrospective analysis comparing screening outcomes before/after implementation of DBT (Hologic Dimensions system) from a single hospital site to national data from Taiwan's National Cancer Registry		9.0%-10.1%	11.4% - 12.2%	17.8% decrease (*p*<.01)				
Powell et al., 2017	FFDM + DBT: 2304 FFDM: 10,477	Retrospective observational data review of images generated with Hologic's Selenia + Dimensions systems		14%	16%	12.5% decrease (*p*=.017)				
Rafferty et al., 2017 (PROSPR consortium centres)	FFDM + DBT: 173,414 FFDM: 278,906	Retrospective, multicentre analysis of images taken using Hologic's Selenia Dimensions system								
Conant et al., 2016	FFDM + DBT: 55,998 FFDM: 142,883 Women aged 40 to 74 years	Retrospective analysis of data from three PROSPR consortium sites (NB mammography system used not stated)		8.7%	10.4%	15.6% decrease (*p*<.0001)				
McDonald et al., 2016	FFDM + DBT: 33,740 FFDM: 10,728 *12079 had one screen* *6293 had two screens* *3023 had three screens*	Retrospective review of mammography metrics from a single site over four years of screening with single reading using Hologic Dimensions system								
Sharpe et al., 2016	FFDM + DBT: 5703 FFDM: 80,149	Prospective study with a retrospective cohort performed at a single site using Hologic Dimensions system for DBT and								

		GE Senographe Essential, 2000D and DS systems								
Wang et al., 2016	FFDM + DBT: 12,444 FFDM: 12,444	Retrospective study of DBT and FFDM images (Hologic Dimensions system) of 65 breast cancers								
Zuckerman et al., 2016	FFDM + DBT: 15,571 DBT + s2DM: 5366	Observational study set in a community screening setting using Hologic Dimensions system	7.1% (statistically significant decrease vs. FFDM + DBT: *p*<.001)	8.8%	NR					
Durand et al., 2015	FFDM + DBT: 8591 FFDM: 9364	Retrospective review which includes CAD using Hologic Selenia and Dimensions systems		7.8%	12.3%	36.6% decrease (*p*<.0001)				
Lourenco et al., 2015	FFDM + DBT: 12,921 (ages 30.9-89.4 years, average age = 54.6 years) FFDM: 12,577 (ages 29.4-90.6 years, average age = 55.3 years)	Retrospective review of two cohorts (DBT alone=2012/13, FFDM=2011/12), single reading with CAD. FFDM performed using GE Senographe series. DBT performed with Hologic Selenia Dimensions system.		6.4%	9.3%	31% decrease(*p*<.00001)		5.94%	8.80%	28.7 decrease per 1000 screens (-35.1 to -22.2)
McDonald et al., 2015	FFDM + DBT (Prevalent): 1859 FFDM + DBT (Incident): 9524 FFDM (Prevalent): 1204 FFDM (Incident): 13,712	Observational study using Hologic Dimensions system for women presenting for mammographic screening at a single institution *NB Data comes from McCarthy et al., 2015*		8.8%	10.4%	22% decrease (*p*=.002)				

Destounis et al., 2014	524 women aged >30 years (mean age 59 years) including women with a history of breast cancer	Retrospective review of images with double reading. FFDM system was Hologic Selenia or Dimensions, GE Senographe Essential or Fuji CRm. DBT system was Hologic Selenia Dimensions.		4.2% (p<0.0001)	11.45%	NR		3.63%	11.07%	74.4 decrease per 1000 screens (-105.6 to -43.1)
Friedewald et al., 2014	FFDM + DBT: 173,663 (ages 52.6-59.7 years, average age = 56.2 years) FFDM: 281,187 (ages 54.4-60.5 years, average age = 57.0 years)	Retrospective review with single reader using data from 13 centres all using Hologic Selenia Dimensions systems		9.1%	10.7%	16.1% decrease (*p*<.001)		8.4%	10.14%	17.4 decrease per 1000 screens (-15.6 to -19.2)
Greenburg et al., 2014	FFDM + DBT: 20,943 FFDM: 38,674 No differences in study arms by age, ethnicity, family history of BC, or prevalence or incidence screening	Retrospective review of mammography outcomes at a multi-site radiology practice using Hologic Selenia or Selenia Dimensions systems		13.6%	16.2%	13.6% decrease (*p*<.0001)				
McCarthy et al., 2014	FFDM + DBT: 15,571 (average age = 56.7 years) FFDDM: 10,728 (average age = 56.9 years)	Observational study using Hologic Dimensions system for women presenting for mammographic screening at a single institution *NB McDonald et al., (2015) reported on baseline/prevalence CDR based on this study*		8.8%	10.4%	16% decrease (*p*<.001)				
Rose et al., 2014	10,878 FFDM+DBT images and 10,878 matched FFDM images	Observational reading study of data before/after DBT implemented		5.5%	8.7%		NR			

Haas et al., 2013	FFDM + DBT: 6100 FFDM alone: 7058	Retrospective analysis using Hologic Selenia and Dimensions systems *NB: No rate is statistically significant*		8.4%	12%	29.7% decrease (p<.01)				
Rose et al., 2013	FFDM + DBT images: 10,878 (average age = 54.5 years) FFDM images: 10,878 (average age = 53.8 years)	Observational reading study of data before/after DBT implemented using Hologic Selenia and Dimensions systems								

APPENDIX B: QUALITY ASSESSMENT FOR EACH INCLUDED STUDY

AMSTAR2 Tool for systematic reviews and meta-analysis

Coop et al., 2016

	AMSTAR2 TOOL QUESTION	Answer	Comment
1	Did the research questions and inclusion criteria for the review include the components of the PICO?	No	Included studies set in both diagnostic and screening settings
2	Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?	Yes	Used PRISMA guidelines
3	Did the review authors explain their selection of the study designs for inclusion in the review?	Yes	
4	Did the review authors use a comprehensive literature search strategy?	Not sure	Only looked at Scopus and Academic One File
5	Was there duplicate study selection and data extraction?	Not sure	
6	Did the review authors provide a list of excluded studies and justify the exclusion?	No	Exclusion criteria were described: not in English, as well as any published before 2005 due to DBT only becoming clinically available after this point. Studies were also excluded if they compared DBT to film-screen mammography, since current screening standards use digital breast mammography.
7	Did the review authors describe the included studies in adequate detail?	No	Included studies were not clearly identified in the text.
8	Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	Not stated	

	AMSTAR2 TOOL QUESTION	Answer	Comment
9	Did the review authors report on the sources of funding for the studies included in the review?	No	
10	If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?	No	No pooled analysis was completed.
11	If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	NA	
12	Did the review authors account for RoB in individual studies when interpreting/discussing the results of the review?	Not clear	RoB is not discussed
13	Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	Yes	
14	If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	NA	
15	Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	No	

Hodgson et al., 2016

	AMSTAR2 TOOL QUESTION	Answer	Comment
1	Did the research questions and inclusion criteria for the review include the components of the PICO?	Yes	
2	Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?	Yes	
3	Did the review authors explain their selection of the study designs for inclusion in the review?	Yes	
4	Did the review authors use a comprehensive literature search strategy?	Yes	
5	Was there duplicate study selection and data extraction?	Not clear	
6	Did the review authors provide a list of excluded studies and justify the exclusion?	Yes	
7	Did the review authors describe the included studies in adequate detail?	Yes	
8	Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	Yes	QUADAS-2 tool used
9	Did the review authors report on the sources of funding for the studies included in the review?	No	
10	If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?	Yes	Some meta-analysis performed where possible
11	If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	Yes	QUADAS-2 tool used

	AMSTAR2 TOOL QUESTION	Answer	Comment
12	Did the review authors account for RoB in individual studies when interpreting/discussing the results of the review?	Yes	
13	Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	Yes	
14	If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	NA	
15	Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	Yes	CoI statement included employment by Novartis Healthcare Pvt. Ltd. (ceased before work commenced on the review) and research funding from Siemens Healthcare for two of the authors.

SIGN criteria for RCTs

Maxwell et al. (2017)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
1.2	The assignment of subjects to treatment groups is randomised.	Yes ☑	Can't say ☐	No ☐
1.3	An adequate concealment method is used.	Yes ☑	Can't say ☐	No ☐
1.4	The design keeps subjects and investigators 'blind' about treatment allocation.	Yes ☐	Can't say ☐	No ☑
1.5	The treatment and control groups are similar at the start of the trial.	Yes ☐	Can't say ☑	No
1.6	The only difference between groups is the treatment under investigation.	Yes ☐	Can't say ☑	No ☐
1.7	All relevant outcomes are measured in a standard, valid and reliable way.	Yes ☑	Can't say ☐	No ☐
1.8	What percentage of the individuals or clusters recruited into each treatment arm of the study dropped out before the study was completed?	9.1% of participants did not complete the second annual screen. No information about which arm these women were in is provided.		
1.9	All the subjects are analysed in the groups to which they were randomly allocated (often referred to as intention to treat analysis).	Yes ☑	Can't say ☐	No ☐
1.10	Where the study is carried out at more than one site, results are comparable for all sites.	Yes ☐	Can't say ☐	No ☑ NA ☐

OVERALL ASSESSMENT OF THE STUDY		
2.1	How well was the study done to minimise bias?	High quality (++)☐ Acceptable (+)☐ Low quality (-) ☑

OVERALL ASSESSMENT OF THE STUDY		
		Unacceptable – reject 0 ☐
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, are you certain that the overall effect is due to the study intervention?	No
2.3	Are the results of this study directly applicable to the patient group targeted by this guideline?	Yes
2.4	**NOTES** *The process for identifying and recruiting participants differed between the two centres. Limited demographic information (aside from age and a global description of cancer risk) about participants is provided so it is not possible to determine if the groups are otherwise similar. There were significant differences in the recall rates by study centre (centre A had significantly higher recall rates); cancer detection rates were comparable by study centre for FFDM but more cancers with detected with DBT + FFDM at centre A. All reported p-values are significantly greater than 0.05 indicating weak evidence for recall results. High loss to follow-up.*	

SIGN criteria for case-control studies (including fully paired trials)

Malmö trial (three studies)

Lang et al. (2016A; 2016B)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The cases and controls are taken from comparable populations.	Yes ☑	Can't say ☐	No ☐
1.3	The same exclusion criteria are used for both cases and controls.	Yes ☑	Can't say ☐	No ☐
1.4	What percentage of each group (cases and controls) participated in the study?	Fully paired		
1.5	Comparison is made between participants and non-participants to establish their similarities or differences.	Yes ☐	Can't say ☑	No ☐
1.6	Cases are clearly defined and differentiated from controls.	Yes ☑	Can't say ☐	No ☐
1.7	It is clearly established that controls are non-cases.	Yes ☑	Can't say ☐	No ☐
ASSESSMENT				
1.8	Measures will have been taken to prevent knowledge of primary exposure influencing case ascertainment.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Exposure status is measured in a standard, valid and reliable way.	Yes ☑	Can't say ☐	No ☐
CONFOUNDING				
1.10	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☑	Can't say ☐	No ☐

STATISTICAL ANALYSIS			
1.11	Confidence intervals are provided.	Yes ☑	No ☐

OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☑ Acceptable (+)☐ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	**NOTES** Prospective, fully paired trial embedded in a population-based screening program in Sweden using one brand of equipment (Siemens). Wo reading arms were used (DBT_{MLO} + DM_{CC} compared to FFDM). The study population has a high number of women participating at a prevalent screen, which may inflate the CDR (either because CDR is higher in this group or because previous images were not available to readers). No data on long-term outcomes was collected as dataset was not large enough or the trial long enough.			

Rosso et al. (2015)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The cases and controls are taken from comparable populations.	Yes ☑	Can't say ☐	No ☐
1.3	The same exclusion criteria are used for both cases and controls.	Yes ☑	Can't say ☐	No ☐
1.4	What percentage of each group (cases and controls) participated in the study?	Fully paired		
1.5	Comparison is made between participants and non-participants to establish their similarities or differences.	Yes ☐	Can't say ☑	No ☐
1.6	Cases are clearly defined and differentiated from controls.	Yes ☑	Can't say ☐	No ☐
1.7	It is clearly established that controls are non-cases.	Yes ☑	Can't say ☐	No ☐
ASSESSMENT				
1.8	Measures will have been taken to prevent knowledge of primary exposure influencing case ascertainment.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Exposure status is measured in a standard, valid and reliable way.	Yes ☑	Can't say ☐	No ☐
CONFOUNDING				
1.10	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☑	Can't say ☐	No ☐
STATISTICAL ANALYSIS				
1.11	Confidence intervals are provided.	Yes ☑	No ☐	

OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☑ Acceptable (+)☐ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	**NOTES** Prospective, fully paired trial embedded in a population-based screening program in Sweden. A range of appropriate statistical techniques used to determine results (used classification and regression tree and binary marginal generalized linear models). Models are limited to co-variates discussed. Radiologist reading experience may also affect the results (generally, participating radiologists were familiar with DBT as a modality).			

OTS trial

Skaane et al. (2014)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The cases and controls are taken from comparable populations.	Yes ☑	Can't say ☐	No ☐
1.3	The same exclusion criteria are used for both cases and controls.	Yes ☑	Can't say ☐	No ☐
1.4	What percentage of each group (cases and controls) participated in the study?	Fully paired		
1.5	Comparison is made between participants and non-participants to establish their similarities or differences.	Yes ☐	Can't say ☑	No ☐
1.6	Cases are clearly defined and differentiated from controls.	Yes ☑	Can't say ☐	No ☐
1.7	It is clearly established that controls are non-cases.	Yes ☑	Can't say ☐	No ☐
ASSESSMENT				
1.8	Measures will have been taken to prevent knowledge of primary exposure influencing case ascertainment.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Exposure status is measured in a standard, valid and reliable way.	Yes ☑	Can't say ☐	No ☐
CONFOUNDING				
1.10	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☑	Can't say ☐	No ☐
STATISTICAL ANALYSIS				

STATISTICAL ANALYSIS			
1.11	Confidence intervals are provided.	Yes ☐	No ☑

OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☑ Acceptable (+)☐ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	**NOTES** Prospective, fully paired trial embedded in a population-based screening program in Norway. *P* values are provided appropriately. The approach to deciding whom to recall may have affected some of the results (as discussed in the text of this literature review). The study population has a high number of women participating at a prevalent screen, which may inflate the CDR (either because CDR is higher in this group or because previous images were not available to readers). Validation of equipment in other settings is needed. No data on long-term outcomes was collected as dataset was not large enough or the trial long enough.			

Skaane et al. (2013A and 2013B)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The cases and controls are taken from comparable populations.	Yes ☑	Can't say ☐	No ☐
1.3	The same exclusion criteria are used for both cases and controls.	Yes ☑	Can't say ☐	No ☐
1.4	What percentage of each group (cases and controls) participated in the study?	Fully paired		
1.5	Comparison is made between participants and non-participants to establish their similarities or differences.	Yes ☐	Can't say ☑	No ☐
1.6	Cases are clearly defined and differentiated from controls.	Yes ☑	Can't say ☐	No ☐
1.7	It is clearly established that controls are non-cases.	Yes ☑	Can't say ☐	No ☐
ASSESSMENT				
1.8	Measures will have been taken to prevent knowledge of primary exposure influencing case ascertainment.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Exposure status is measured in a standard, valid and reliable way.	Yes ☑	Can't say ☐	No ☐
CONFOUNDING				
1.10	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☑	Can't say ☐	No ☐
STATISTICAL ANALYSIS				
1.11	Confidence intervals are provided.	Yes ☐	No ☑	

OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☑ Acceptable (+)☐ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	**NOTES** Prospective, fully paired trial embedded in a population-based screening program in Norway. Four reading arms were used. *P* values are provided appropriately. Reading and arbitration method could affect recall rates and the use of computer aided detection could have influenced results by decreasing actual recall rates. Readers did not interpret the same number of images, but this was adjusted for in the analysis. No data on long-term outcomes was collected as dataset was not large enough or the trial long enough.			

STORM

Ciatto et al. (2013: main study)

Other studies using data from STORM included Bernardi et al., (2014), Houssami et al., (2014) and Caumo et al., (2013).

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The cases and controls are taken from comparable populations.	Yes ☑	Can't say ☐	No ☐
1.3	The same exclusion criteria are used for both cases and controls.	Yes ☑	Can't say ☐	No ☐
1.4	What percentage of each group (cases and controls) participated in the study?	Sequential reading by radiologists		
1.5	Comparison is made between participants and non-participants to establish their similarities or differences.	Yes ☐	Can't say ☑	No ☐
1.6	Cases are clearly defined and differentiated from controls.	Yes ☑	Can't say ☐	No ☐
1.7	It is clearly established that controls are non-cases.	Yes ☑	Can't say ☐	No ☐
ASSESSMENT				
1.8	Measures will have been taken to prevent knowledge of primary exposure influencing case ascertainment.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Exposure status is measured in a standard, valid and reliable way.	Yes ☑	Can't say ☐	No ☐
CONFOUNDING				
1.10	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☑	Can't say ☐	No ☐

STATISTICAL ANALYSIS			
1.11	Confidence intervals are provided.	Yes ☑	No ☐

OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☑ Acceptable (+)☐ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	**NOTES** Prospective trial using sequential reading by radiologists in a single population attending for screening (Verona and Trento, Italy). FFDM first, then FFDM + DBT. Confounders associated with reader bias are discussed (eg, low numbers of screens read by individual radiologists which could affect interpretation efficiency and limited statistical analysis, reading protocol could have affected threshold to recall especially as FFDM results were read before DBT results and there could have been recall bias based on review of FFDM images at a later time as well). No data on long-term outcomes was collected as dataset was not large enough or the trial long enough.			

STORM-2

Bernardi et al. (2016)

Other studies reporting on STORM-2 data included Bernardi and Houssami (2017B), and Houssami et al., (2017).

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The cases and controls are taken from comparable populations.	Yes ☑	Can't say ☐	No ☐
1.3	The same exclusion criteria are used for both cases and controls.	Yes ☑	Can't say ☐	No ☐
1.4	What percentage of each group (cases and controls) participated in the study?	Fully paired		
1.5	Comparison is made between participants and non-participants to establish their similarities or differences.	Yes ☐	Can't say ☑	No ☐
1.6	Cases are clearly defined and differentiated from controls.	Yes ☑	Can't say ☐	No ☐
1.7	It is clearly established that controls are non-cases.	Yes ☑	Can't say ☐	No ☐
ASSESSMENT				
1.8	Measures will have been taken to prevent knowledge of primary exposure influencing case ascertainment.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Exposure status is measured in a standard, valid and reliable way.	Yes ☑	Can't say ☐	No ☐
CONFOUNDING				
1.10	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☑	Can't say ☐	No ☐

STATISTICAL ANALYSIS			
1.11	Confidence intervals are provided.	Yes ☑	No ☐

OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☑ Acceptable (+)☐ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	**NOTES** Prospective trial in a single population attending for screening (Trento, Italy). Reading was parallel, sequential and double. Reading arms effectively resulted in four reads so CDR may be inflated if data from all arms is presented (however, the authors only presented data from the double reading strategy). Reading strategy may also have conflated the recall rate. This study was also readers' first experience with s2DM. No data on long-term outcomes was collected as dataset was not large enough or the trial long enough.			

SIGN criteria for other studies

Aujero et al.

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The cases and controls are taken from comparable populations.	Yes ☑	Can't say ☐	No ☐
1.3	The same exclusion criteria are used for both cases and controls.	Yes ☐	Can't say ☑	No ☐
1.4	What percentage of each group (cases and controls) participated in the study?	Cases: 16,173 s2DM+DBT (20.5%) Controls: 30,561 FFDM+DBT (38.8%) 32,076 FFDM alone (40.7%)		
1.5	Comparison is made between participants and non-participants to establish their similarities or differences.	Yes ☐	Can't say ☑	No ☐
1.6	Cases are clearly defined and differentiated from controls.	Yes ☑	Can't say ☐	No ☐
1.7	It is clearly established that controls are non-cases.	Yes ☑	Can't say ☐	No ☐
ASSESSMENT				
1.8	Measures will have been taken to prevent knowledge of primary exposure influencing case ascertainment.	Yes ☐	Can't say ☑	No ☐ NA ☐
1.9	Exposure status is measured in a standard, valid and reliable way.	Yes☑	Can't say ☐	No ☐
CONFOUNDING				
1.10	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☐	No ☑

STATISTICAL ANALYSIS			
1.11	Confidence intervals are provided.	Yes ☑	No ☐

OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective study with missing data for some included results. Reader experience with s2DM may have also influenced the results. No data on long-term outcomes was collected as dataset was not large enough or the trial long enough.			

Bernardi et al. (2012)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☑	No ☐ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.	Not stated		
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☑	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomization.			

Carbonaro et al. (2016)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.	15%: 41 out of 173 patients were lost to follow-up (negative triple assessment)		
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☐
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☐	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☐	No ☑
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomization. All cases were recalled for some reason (so results are no representative of a general screening population). Some patients lost to follow-up. Likely to over-estimate in reduced recall rate because of the study inclusion criteria (women recalled from screening). All recalls were considered reviewed by double reading when some may have been recalled with single reading.			

Conant et al. (2016)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☐	Can't say ☑	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.	Not reported		
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☑	Can't say ☐	No ☐ NA ☐
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☑	No ☐ NA ☐
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☑	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☐	No ☑
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomization. Multisite analysis with limited information provided about the differences between populations aside from noting that there were differences in ethnicity, age and potential breast cancer risk factors. Different sites had different levels of experience with DBT. No information about long-term outcomes was provided.			

Dang et al. (2014)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.	Not stated		
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No☑ NA ☐
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☑	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☐	No ☑	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☐	Can't say ☑	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomization. Interpretation from radiologists who were experienced with DBT (which may have increased interpretation times). It is not clear whether the volume of results read represents what would happen in a clinical environment (it may be higher).			

Destounis et al. (2014)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☑	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☐	No ☑
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☐	No☑	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☐ Unacceptable – reject 0 ☑		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☐	Can't say ☑	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☐	No ☑	
2.4	Retrospective analysis with no randomization. Patient bias is likely. Patient age range is 30-90 years. A small fee was initially charged at the beginning of the study, with an increase in the number of patients choosing to have the exam after the fee was stopped. Patients were made aware of the increased radiation dose associated with the exam. Patients with high risk factors may have preselected themselves to have their exam with new technology.			

Durand et al. (2014)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.	Data not available		
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☑	Can't say ☐	No ☐ NA ☐
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☑	No ☐ NA ☐
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☑	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☑ Acceptable (+)☐ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomization. No follow-up data about longer-term clinical health outcomes (one screening examination only).			

Freer et al. (2017)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.	Data not available		
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☑	No ☐ NA ☐
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☑	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☐	No ☑
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☑ Acceptable (+)☐ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomization. No assessment of potential confounders across the two time periods completed. DBT was not implemented at the same time across different sites. No long-term outcomes data defined.			

Friedewald et al. (2014)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☑	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomisation. Did not stratify by age. DBT introduction at different sites was not uniform due to budgeting constraints at most sites.			

Greenburg et al. (2014)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☑	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☐	No ☑
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomisation. Patients were offered DBT at no additional charge (1147 patients). After a point, DBT was offered at an additional $50 fee. It is unclear how this was taken into the analysis.			

Gur et al. (2012)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☑	No ☐ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☑	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomisation.			

Lourenco et al. (2015)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☑ Acceptable (+)☐ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomisation.			

McCarthy et al. (2014)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☐	Can't say ☐	No ☑
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☑	Can't say ☐	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomisation. DBT was not implemented in at a single time but cohorts are similar. Multivariate analysis performed to address population variation. Not powered to detection incremental changes in CDR or long-term clinical outcomes.			

McDonald et al. (2015)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☑	Can't say ☐	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomisation. Baseline screening categories did not include only women having a prevalent screen.			

McDonald et al. (2016)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☐	No ☑
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomization but covered a screening population. No information about risk characteristics was provided. High risk patients did not receive DBT (had a diagnostic assessment) so study population may not represent the same population covered in a screening program. No data about financial cost-effectiveness or long-term clinical outcomes was provided.			

Pan et al. (2017)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☑	No ☐ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.	Not stated		
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☐	No ☑	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomization but covered a screening population. Information about the study population was hard to ascertain.			

Powell et al. (2017)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☑	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomization. Women with risk factors for breast cancer were more likely to receive DBT, leading to bias in assignation. Self-selection bias is also present.			

Olgar et al. (2012)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4				

Rafferty et al. (2017)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☑	Can't say ☐	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective analysis with no randomization. Follow-up data is not available meaning that no long-term outcomes can be assessed. No information available about prevalent or incident screening.			

Rafferty et al. (2014)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☑	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☑	Can't say ☐	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	

Rodriguez-Ruiz et al. (2017)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective, multicenter analysis with a high rate of participants who were recalled from screening with FFDM, which means the CDR may be higher than expected with DBT compared to FFDM.			

Rose et al. (2013)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☐	No ☑	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective observational study that may be affected by self-selection bias from participants. No long-term data is available for this study population.			

Rose et al. (2014)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☐	No ☑	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective reader study that could be affected by inter-reader variability because of the way that screening examinations were allocated. FFDM readings may be affected by knowing DBT results first.			

Sharpe et al. (2016)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Prospective investigation of a retrospective cohort but some patients requested reallocation from the group they were assigned to.			

Shin et al. (2015)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	**NOTES**			

Starikov et al. (2015)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☐	No ☑	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Comparative and retrospective observer performance evaluation assessing an enriched data set which means readers are likely to recall more cases that would be expected under a normal screening protocol. Limited numbers of study participants had very fatty breasts (limiting conclusions on density findings).			

Sumkin et al. (2015)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☑	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☑	No ☐ NA ☐
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☐	No ☑	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Has a high clinical baseline for recall and large inter-reader variability. Participants could self-select to participate in this study.			

Zuckerman et al. (2016)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4				

Zuley et al. (2014)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☑	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☑	No ☐	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	**NOTES**			

Zuley et al. (2010)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☑ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.			
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☑	Can't say ☐	No ☐ NA ☐
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☐	No ☑	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	**NOTES**			

SIGN criteria for cohort studies

Abdullah Suhaimi SA et al. (2015)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☑	No ☐	NA ☐
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.	N/A		
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☐	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☐	Can't say ☐	No ☐ NA ☑
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☑	Can't say ☐	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☐	No ☑	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☑	Can't say ☐	No ☐
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	**NOTES** 130 Malaysian women aged 40-69 years. Questionnaire used to measure the state anxiety level after standard and reduced compression mammography.			

Wang et al. (2016)

INTERNAL VALIDITY				
1.1	The study addresses an appropriate and clearly focused question.	Yes ☑	Can't say ☐	No ☐
SELECTION OF SUBJECTS				
1.2	The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation.	Yes ☑	Can't say ☐	No ☐ NA ☐
1.3	The study indicates how many of the people asked to take part did so, in each of the groups being studied.	Yes ☐	No ☐	NA ☑
1.4	The likelihood that some eligible subjects might have the outcome at the time of enrolment is assessed and taken into account in the analysis.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.5	What percentage of individuals or clusters recruited into each arm of the study dropped out before the study was completed.	N/A		
1.6	Comparison is made between full participants and those lost to follow up, by exposure status.	Yes ☐	Can't say ☐	No ☐ NA ☑
ASSESSMENT				
1.7	The outcomes are clearly defined.	Yes ☑	Can't say ☐	No ☐
1.8	The assessment of outcome is made blind to exposure status. If the study is retrospective this may not be applicable.	Yes ☐	Can't say ☐	No ☐ NA ☑
1.9	Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment of outcome.	Yes ☐	Can't say ☑	No ☐
1.10	The method of assessment of exposure is reliable	Yes ☑	Can't say ☐	No ☐
1.11	Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable.	Yes ☑	Can't say ☐	No ☐

ASSESSMENT				
				NA ☐
1.12	Exposure level or prognostic factor is assessed more than once.	Yes ☐	Can't say ☐	No ☐ NA ☑
CONFOUNDING				
1.13	The main potential confounders are identified and taken into account in the design and analysis.	Yes ☐	Can't say ☑	No ☐
STATISTICAL ANALYSIS				
1.13	Have confidence intervals been provided?	Yes ☐	No ☑	
OVERALL ASSESSMENT OF THE STUDY				
2.1	How well was the study done to minimise the risk of bias or confounding?	High quality (++)☐ Acceptable (+)☑ Unacceptable – reject 0 ☐		
2.2	Taking into account clinical considerations, your evaluation of the methodology used, and the statistical power of the study, do you think there is clear evidence of an association between exposure and outcome?	Yes ☐	Can't say ☐	No ☑
2.3	Are the results of this study directly applicable to the patient group targeted in this guideline?	Yes ☑	No ☐	
2.4	Retrospective review of pathology and histologic findings in a diagnostic population at a single site (all breast cancers diagnosed). Small number of cancers diagnosis and study has limited ability to report on longer-term clinical outcomes. To be confirmed as mammographically occult, all five radiologists needed to agree.			