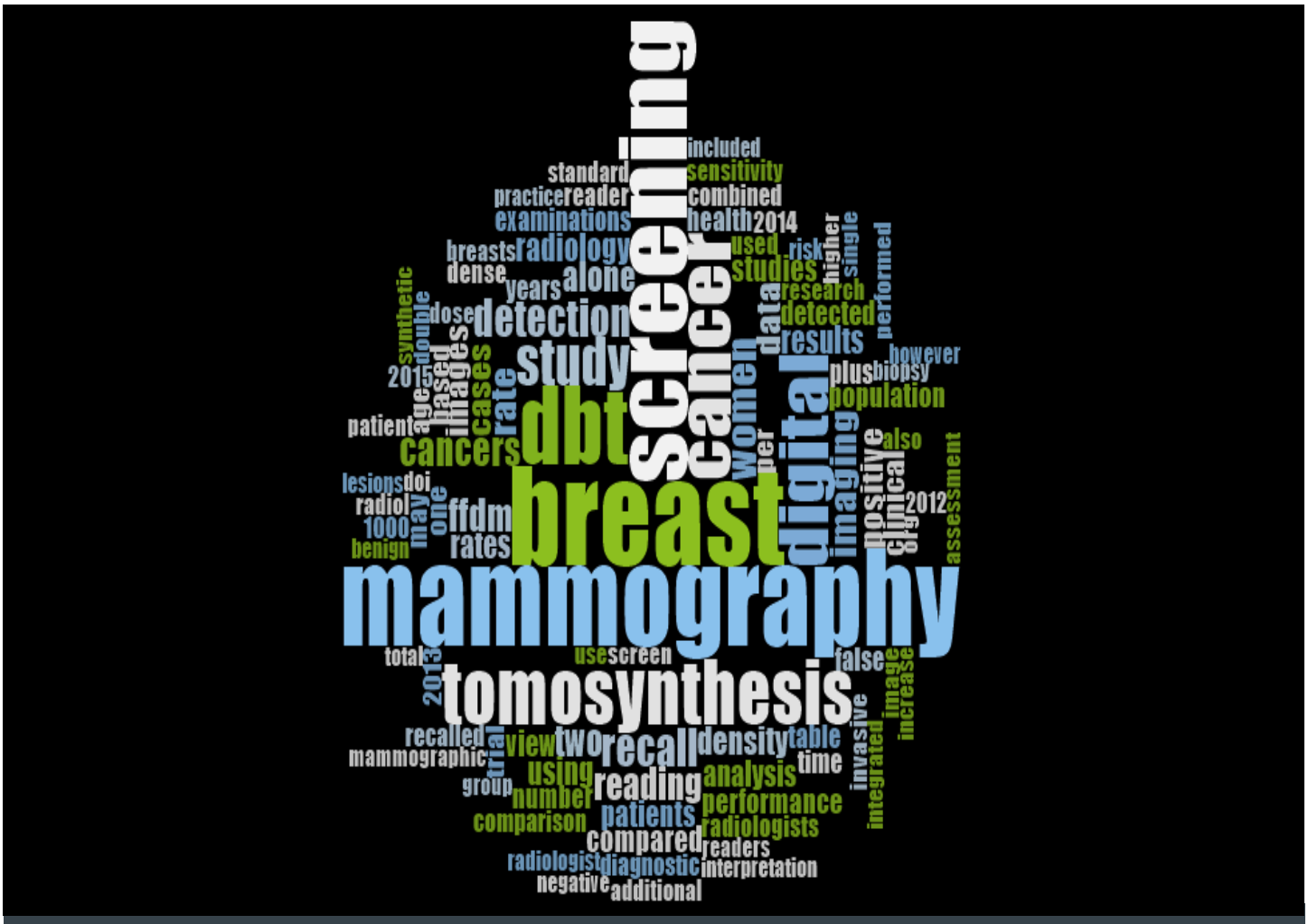


Digital breast tomosynthesis

An update of *Allen + Clarke*'s literature reviews on the use of tomosynthesis in the BreastScreen Australia program

Final report: 8 May 2020

ALLEN+CLARKE

Document status:	Final report
Version and date:	V3, 18 June 2020
Author(s):	Anna Gribble, Sarah Renals
Filing Location:	W:\Department of Health and Ageing Australia\Breastscreen Australia 2017\Deliverables\Tomosynthesis\Tomo in screening\2019 update
Verification that QA changes made:	Anna Gribble
Proofread:	Rob Smith, Sarah Renals
Formatting:	Anna Gribble
Final QA check and approved for release:	Anna Gribble

Allen + Clarke has been independently certified as compliant with ISO9001:2015 Quality Management Systems

Telarc. Registered TM
Quality ISO 9001

CONTENTS

ACRONYMS

95%CI	95% confidence interval
AD	Architectural distortion
BIRADS	Breast Imaging Reporting and Data System
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)
DCIS	Ductal carcinoma in situ
DICOM	Digital Imaging and Communications
DM	Digital mammography
DM_{CC}	One view digital mammography (craniocaudal view)
FFDM	Full-field digital mammography (also known as two-view digital mammography)
HER-2	Human epidermal growth factor receptor 2
IDC	Invasive ductal carcinoma
ILC	Invasive lobular carcinoma
JAFROC	Jackknife free-response receiver operating characteristic
MGD	Mean glandular dose
mGy	Milligray
MLO	Mediolateral oblique (view)
OR	Odds ratio
OR+	Oestrogen positive receptor
OTS	Oslo Tomosynthesis in Screening trial
PPV	Positive predictive value
PR+	Progesterone positive receptor
PROSPR	Population-based Research to Optimize the Screening Process consortium
QALY	Quality-adjusted life year
s2DM	Synthesised two-view digital mammography
STORM	Screening with Tomosynthesis or Regular Mammography trial
RR	Relative risk
VDG	Volpara density grade

GUIDANCE ON HOW TO READ THIS REPORT

This report is a narrative literature review. It builds on two previous reports completed by *Allen + Clarke*: one narrative literature review on the role of tomosynthesis in screening (which covered literature published on tomosynthesis between 1 January 2010 and 31 December 2017), and another on the role of tomosynthesis in the assessment clinic (which covered literature published between 1 January 2010 and 30 May 2018). Both reports are available at http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/dbt.

This narrative literature review focuses on DBT in the screening environment. It contains two main parts:

1. The *Key Findings* section provides a summary of the findings of this review. 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 also answer the research questions in a summary analysis in this section.
2. The main report provides detailed findings 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, positive predictive value, recall rate, etc.), we have presented the information by clinical outcome or performance metric rather than by study. Information is presented in a hierarchy of evidence and evidence tables are included in each section. For ease of reference, we start each outcome/performance measure section with a short summary describing findings from the earlier two literature reviews.

Appendix A includes the quality assessment tables (based on AMSTAR2 and the Scottish Intercollegiate Guidelines Network tools) for included systematic reviews and randomized controlled trials (RCT).

KEY FINDINGS

The BreastScreen Australia (BSA) program currently uses bilateral full-field digital mammography (FFDM) as the "gold standard" screening test for the for the early detection of breast cancer in asymptomatic women aged 40-74 years.

The Department of Health (Australia) has previously commissioned *Allen + Clarke* to undertake two literature reviews (not systematic reviews) 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. One review focused on the role of DBT as a primary screening test (looking at literature published to December 2017); the second reviewed the role of DBT in the assessment of screen-detected abnormalities (literature published to June 2018). The role of DBT in the screening environment is an area of active research, with several large trials underway and pilot evaluations reporting baseline, interim or initial results (including results from the Maroondah pilot). It is therefore timely to complete an updated analysis of the literature investigating the role of DBT in the screening environment.

Methodology

Allen + Clarke completed a systematic search of the Ovid Medline 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/S) criteria. The same research questions included in the 2018 literature review on the role of DBT in screening were used.

In this update, a total of 41 articles met the inclusion criteria:

- Two systematic reviews covering immediate screening outcomes
- Five narrative literature reviews or editorial commentary by experienced researchers
- Four primary research papers from two RCTs
- Nine prospective studies embedded in a European population-based screening program and three sub-studies from the STORM-2 trial
- Two prospective cohort studies embedded in a European population-based screening program that used a historical cohort
- One prospective study embedded in an American screening program
- Three observer performance studies
- 11 retrospective analyses, and
- One cost-effectiveness study.

Results

Most research focused on shorter-term performance measures like cancer detection rate (CDR), recall rate and cancer type rather than longer-term screening outcomes such as mortality reduction or a reduction in interval cancers. Uncertainty about the effectiveness of DBT as the primary test in a population-based screening program remains (and is likely to remain until larger

long-term studies report results). That said, digital breast tomosynthesis plus synthesised 2D mammography (DBT+s2DM) has been implemented as the primary screening test in several European and American screening programs, with initial results reporting improvement in short-term outcomes for program sensitivity.

Sensitivity

Overall CDR

The findings add further depth to *Allen + Clarke*'s 2018 literature reviews but some mixed results were reported in the To-Be-1 RCT (the first large RCT of DBT). Pooled evidence from papers published before December 2017 (involving more than one million women) reported a statistically significant increase in CDR, with an incremental increase in detection of 1.6 cancers per 1000 screening examinations. Higher increases were reported in European population-based programs with similar policy settings to the BreastScreen Australia program (i.e., pooled analysis indicated 2.4 more cancers were detected when DBT was used).

Studies published since 1 January 2018 are generally consistent with the pooled analysis (that use of DBT significantly increases CDR and that DBT+s2DM is not inferior to FFDM+DBT or FFDM alone), with one important exception. The To-Be-1 RCT found no statistically significant increase in CDR with the use of DBT+s2DM. Possible explanations for this difference were that all previous images were available to readers, which may have favoured FFDM imaging (and which probably better reflects a 'real world' screening environment). There may be other study differences to which require further unpicking. That said, evidence is still strong that use of DBT increases cancer detection.

Increases in CDR were reported in all other studies (including large, ongoing robust prospective trials and robust retrospective analyses) for different combinations of screening strategy including FFDM+DBT, DBT+s2DM, and DBT_{MLO} compared to FFDM (the Malmö trial). Evidence also suggests that DBT+s2DM is superior to FFDM alone for CDR and is not inferior to FFDM+DBT. For this reason, where screening programs have chosen to implement DBT, they have chosen DBT+s2DM rather than dual acquisition. Other imaging protocols were also explored (including DBT_{MLO} or single reading of DBT), again favouring the use of DBT in terms of increasing cancer detection.

Evidence is also emerging that DBT increases CDR in women with more dense breasts and older women. Generally, evidence suggested that use of DBT results in increased CDR for all breast densities (i.e., all women) but some evidence is emerging that the increase may be greater for women whose breasts are more dense. While breast density is not reported in the BSA program, considering results of stratification analysis by density is useful in considering whether there are some population groups for whom DBT may be more beneficial.

Interval cancer rate

Interval cancer rate is an important long-term screening outcome, especially as interval cancers are often more advanced and can be associated with higher morbidity and mortality compared to screen-detected cancers. Several pilot evaluation or paired studies have reported on interval cancer rates; however, reported results have generally come from studies that have not been powered to calculate interval cancer rates or which have varying screen intervals (as per the

program's policy settings). Other studies like the Reggio Emilia RCT are powered to detect interval cancers but results have not yet been reported.

Current data is indicative only. Adequate data evaluating interval cancer rate by imaging modality is not yet available. Reported results vary and no study has reported a statistically significant reduction in interval cancers with DBT. Few studies compared interval cancer characteristics by type, grade, node status and size, which would be helpful to assess whether DBT detects more aggressive cancers earlier. Only the OTS trial has reported this and found no significant differences in the types of cancers detected between pre- and post- screening with DBT. This suggests that DBT may not influence the detection of more aggressive cancers earlier; however, at this stage, we do not know the impact of DBT on decreasing interval cancer rates or whether interval cancers differ in terms of pathology/histology from cancers detected within a screening round or with interval cancers from previous screening rounds. Determining DBT's influence on interval cancers remains one of the key areas to explore in future research. The screening interval (i.e., annual or biennial) and the way that interval cancer rate is defined and reported in studies is important as this could also account for variability within study results. Pooled analysis work, which will include data from the main prospective paired studies set in population-based screening programs, is underway and due to report in 2020.

Relative sensitivity

All studies reported increases in sensitivity when DBT was used but non-significant results or wide confidence intervals were also noted. Increases were observed with all reading protocols (i.e., FFDM+DBT, DBT+s2DM). When increases in sensitivity were observed, the increase ranged from less than one percent more to over 16 percent more. Larger increases in sensitivity were observed in the screening programs most similar to the BSA program (double read, biennial screening).

Tumour characteristics

Overall, use of DBT appears to detect more invasive disease than FFDM (including when DBT is used in combination with FFDM or s2DM or if DBT_{MLO} is used alone). Evidence describing whether DBT detects more DCIS is inconsistent: most study results achieving statistical significance suggested that more DCIS is detected with DBT, but more studies suggest there is no increase or less detection with DCIS. More research is needed to fully unpick these results. It is also clear that DBT+s2DM does not result in less detection of either invasive cancers or DCIS (i.e., s2DM can probably be used instead of FFDM, reducing radiation dose).

Information relating to molecular subtype, grade, size and lymph node status is also important in assessing whether cancers detected are clinically significant. While data was grouped in different bands making it difficult to fully assess this relationship, there is sufficient evidence to suggest that cancers detected with DBT are likely to be slightly smaller, lower grade and node-negative (i.e., potentially earlier stage cancer) than those detected with FFDM. Most studies reported no or only small differences in the proportion of node negative and node positive disease, regardless of the imaging used (i.e., DBT+s2DM, FFDM+DBT or DBT_{MLO}). If there was a difference, it usually favoured the DBT imaging arm (i.e., DBT detected more node-negative disease). Mixed results were also presented for grade, with some studies suggesting that DBT detected more low-grade cancers compared to those detected with FFDM.

Generally, it appears that the cancers detected by DBT are similar in terms of histological type to those detected with FFDM (although more cancers presenting as spiculated masses and architectural distortion are detected with DBT). Across all studies (regardless of the way DBT was used), increases in invasive ductal carcinoma (IDC) were usually observed with the use of DBT. Mixed results were presented for invasive lobular carcinoma (ILC), with some studies demonstrating a statistically significant increase in ILC detection with FFDM, but others noting an increase in ILC detection when DBT was used (including results from the Malmö trial). Mixed results were also presented about molecular subtype. Some studies (including data from the Norway BreastScreen Program and baseline data from an Italian RCT) showed an increase in cancers with a molecular profile more favourable to good prognosis (i.e., more Luminal A cancers, more Luminal B Her-2 negative cancers and more cancers with Ki67%≤20). Others, including data from the To-Be-1 RCT and the Malmö trial showed no significant differences in molecular subtype or receptor status.

Some of these differences may be due to insufficient powering, descriptive rather than comparative analysis, or the way that cancers are grouped according to the imaging in which they were detected. There is not yet sufficient evidence to determine if additional cancers detected with DBT alone are more aggressive 'killing' cancers, if they are comparable to the types of cancers detected with FFDM (but more are detected) or if additional cancers detected may add to the overdiagnosis burden; however, there is accumulating evidence that use of DBT does not contribute to an increase in detection of DCIS (although it could still contribute to overdiagnosis through the detection of less aggressive cancers like Luminal A cancer).

Radiological presentation

Only a small number of studies reported information on radiological presentation by different imaging protocol. Most confirmed previously known information: use of DBT detects more cancers presenting as architectural distortion than FFDM, masses are better defined in a DBT image and therefore DBT detects more cancers presenting as spiculated and circumscribed masses, and that FFDM depicts microcalcifications very well but s2DM is not inferior.

Specificity

Recall rate

The effect of DBT on recall rate varies as does its importance: a decrease in recall rate with DBT may be appropriate in a program which has high recall rates overall but may be less important where rates are already appropriately low.

Two recent systematic reviews indicated mixed results between paired and unpaired studies. In studies where recall rate is quite low already (often those using double reading plus some form of arbitration/consensus), recall rates with DBT were usually a little higher with DBT modalities. The pooled analysis suggested that overall there was a 1.8 percent decrease in recall with FFDM+DBT compared to FFDM alone. In studies where single reading was the norm, recall rates may decline more with the use of DBT. Many studies also observed a learning curve effect, with recall rates for DBT declining in time as readers learnt some of the subtle differences in presentation of cancers on DBT (for example, presentation as architectural distortion). That said, observed recall rates may still have been higher with DBT even after improvements.

Of particular relevance to the BSA were the findings of the Maroondah pilot trial, which reported an increase in recall for DBT+s2DM compared to FFDM alone; however, the overall increase in recall was not extremely large (4.2 recalls per 1000 women screened with DBT+s2DM compared to 3.0 recalls per 1000 women screened with FFDM). The authors proposed that DBT recall rates may decline in Australia as screen readers become more experienced with the technique and previous rounds of DBT images are made available (which is the current status of FFDM).

False positive recall rate

Inconsistent results regarding the influence of DBT on false positive recall rates were reported. Of the six studies that reported results, two prospective studies reported higher false positive recall rates for DBT+s2DM compared to FFDM and four studies reported lower rates of false positive recall. Studies that recorded increases in false positive recall rates for DBT were the STORM-2 trial and Malmö trial. The STORM-2 trial did not employ a means of arbitration for discordant screen readings which may have influenced the result (i.e., a lower false positive recall rate could have been achieved if third party consensus was used). Additionally, other factors that could have impacted higher false positive recall rates in this study suggested by the authors were learning effect in relation to adjusting to new imaging protocols that produce different images and the process of sequential screen-reading at the DBT phase which allowed within participant comparison. Other studies also noted a learning effect in relation to reader lack of familiarity with $DBT_{MLO.}$ Another factors to consider is the lack of availability of prior DBT screens that could be compared unlike FFDM. The inclusion of DBT or DBT+2SDM in population breast screening poses a difficult question in relation to the potential trade-off between increasing CDR but also increasing false positives.

Relative specificity

Studies reported an increase in relative specificity when DBT was used. Most studies reported relative specificity as greater than 91 percent for DBT imaging, and greater than 88 percent for FFDM imaging. Increases were generally very small, ranging from a 0.5 percent increase to 2.4 percent (which indicates FFDM's good performance for this metric, as well as DBT's superior performance). Only one of these studies (data from the OTS trial) was set in a screening program with similar parameters to the BSA program. The study which did not see improved specificity adopted DBT_{MLO} and suggested that an improved specificity would likely be seen with the increase of a second DBT view.

Positive predictive value

There is strong, consistent evidence that use of DBT increases PPV_1 (the percentage of breast cancer cases detected among recalled women) (i.e., DBT is a more accurate test compared to FFM alone. Usually, studies reported that PPV_1 approximately doubled but RCT evidence suggested a slightly lower increase (about 30 percent). PPV_3 (the percentage of breast cancer detected from needle biopsy after recall) also increased significantly with the use of DBT.

Safety

Radiation dose varies with the image technique process used (DBT+s2DM, FFDM alone or FFDM+DBT), the number of and type of views, the use of automatic exposure control, positioning, breast size and composition, and by DBT system used. Most of the literature regarding mean glandular dose (MGD) published since December 2017 reported on MGD for DBT+s2DM

compared to FFDM alone, although some new data on dual acquisition (FFDM+DBT) was also reported.

Mixed results were presented on MGD, including within imaging protocols. For DBT+s2DM compared to FFDM alone, most studies reported that MGD was higher when DBT+s2DM was used although the To-Be-1 RCT, the only study to utilise GE SenoClaire units, reported no significant difference in MGD between DBT+s2DM and FFDM alone. All the other studies reported increases with this DBT modality. Some increases were small (from about 30 percent more) but an increase in MGD of almost double was reported in the Maroondah pilot. The reasons for these differences are not entirely clear although some differences may be due to imaging systems or the use of automated dose (as opposed to radiologist setting). 'Real world' studies all reported a higher per-view MGD when DBT was used compared to FFDM alone (which is consistent with some of the previous technical evaluation findings). Further investigation into the use of different imaging units (Hologic Selenia Dimensions, Siemens Mammomat Inspirations and GE SenoClaire units), software (Volpara Solutions and Quantra™) and protocols in relation to the impact on radiation dose is needed as some studies have started to suggest potential inaccuracies within breast density and radiation dose estimation that could be impacting research results and potentially women being screened.

Implementation

Much of the published literature focused on DBT's sensitivity, specificity and safety with the studies reporting on the nature of the association between DBT (either alone, as an adjunct to FFDM, or with s2DM) and specific clinical outcomes and short-term performance metrics. There is growing confidence that as a screening strategy DBT could enhance a screening program with several programs moving to pilot DBT as the screening test (eg, the Trento and Verona screening programs).

Image acquisition

An increase in image acquisition time could increase the discomfort felt by women when participating in breast screening. Additionally, an increase in time could impact high-throughput clinic workflow. Currently available literature is surprisingly sparse, and certainly insufficient to determine the impact of DBT on image acquisition time. Only one primary study (the To-Be-1 RCT) reported findings on image acquisition time. Results from this RCT were based on the time the woman entered the exam room until the time she left. Results reported from the To-Be-1 RCT stated that DBT+s2DM took 54 sec longer than FFDM (DBT+s2DM imaging took 5 min 24 sec compared to four min 19 sec with FFDM). While the definition of image acquisition differed to other previous studies (being the time in the exam room, rather than the machine settings), the trend appears to remain the same: DBT image acquisition takes longer than FFDM. One explanation presented by the To-Be-1 RCT is that additional time for DBT may have been impacted by explaining to the woman the new technology. We do not know if the actual image acquisition time was much longer (one second is not much more, as reported in other studies). There is also a suggestion that newer machines have the potential to reduce the difference in DBT capture time (observed in early commentary from Moger et al. (2019) in relation to the extension the To-Be-1 RCT (To-Be-2). The Maroondah pilot in Australia may also provide more insight into these issues in the future but it has not yet reported any data on image acquisition.

Reader performance

Few studies described performance metrics related to reader performance and / or undertook stratified analysis of performance across individual radiologists or between groups of radiologists with differing levels of experience (although most studies involved readers with a range of experience in breast imaging and some noted the level of experience in interpreting DBT and s2DM images). Some studies noted pre-trial training in DBT and s2DM screen reading; however, most studies also recognised that this was either insufficient or as in the case of the Malmö trial, not sufficient in a 'real world' setting. It is difficult to determine the extent to which training and prior experience made a difference across the studies that mentioned it. Many studies noted that there is a likely presence of a learning curve, meaning that recall rates tend to drop over time as readers become more familiar with the different type of image that DBT and DBT+s2DM provides. Some studies reported that the learning curve was very short on average with sustained improved screening performance soon after DBT adoption; however, changes in performance with experience may vary across radiologists.

The STORM-2 trial was one of a few studies that provided individual performance data, which indicated that while cancer detection rates across the team of radiologists improved with the addition of DBT, so did the false positive recall rate. Availability of prior DBT images was also considered to be influential to reader performance.

Further investigation is warranted to determine if the experience and training in DBT/s2DM screen reading and the availability of prior images is significantly impacting interpretation times, recall rates, false positive recall rates and cancer detection rates.

Interpretation time

Studies observed significant increases in interpretation time for all DBT imaging modalities, with a doubling of reading time being the most common (which is similar to earlier reported increases). For DBT+s2DM compared to FFDM, the To-Be-1 RCT reported an initial reading time of 1 min 6 sec with DBT+s2DM compared to 39 sec with FFDM. The same was found in the Reggio Emilia RCT: interpretation time for FFDM+DBT was 56 sec compared with 34 sec for FFDM alone. Only one study reported almost no difference in interpretation time between DBT alone and DBT+FFDM if the result was positive; however, no specific data was provided so it is not possible to verify this. Increased reading time will have workflow and reader/radiologist resourcing implications. Some authors suggested that this additional time may be acceptable if DBT delivers less need for consensus (the number of screens requiring arbitration), lower false positive recall rates and higher CDR rates. There continues to be a need for more research into these variables and also with regard to the impact of DBT training and experience on interpretation time.

There was emerging evidence that the addition of computer-aided detection (CAD) to DBT may reduce interpretation time compared with DBT alone; however, this was only explored in one study and there was no comparison to FFDM. One small cancer enriched study provided positive results for the addition of CAD to DBT finding a significant reduction in reading time without a loss of diagnostic performance compared to DBT alone. The effect of reduced reading time was consistently shown by novice and experienced readers (Chae et al., 2019). The benefit of adding CAD to DBT+s2DM may have the potential to reduce reading time; however, limited research has been conducted in this area and needs further exploration. For an area of research as complex as computer aided cancer detection, many more studies will be needed before recommendations can be made regarding its future utility in population-based breast screening programs.

Cost-effectiveness

Two studies provided costing analysis (a Norwegian RCT and a US retrospective modelling analysis). Both indicated that DBT (either instead of or in addition to FFDM) required significant additional costs on a micro and macro level and in both the short- and long-term. They estimated that increased cost would continue to be the case irrespective of more favorable scenarios such as reduction in connectivity and data storage costs, reduction in the price of DBT capable equipment, reduction in DBT reading time, increased DBT sensitivity, and reduced charges for of DBT screening procedures. No Australian cost data was reported in the first (and only) paper from the Maroondah pilot.

While there is interest in exploring the incremental costs of DBT (+/- s2DM) compared to FFDM, there continues to be a paucity of studies investigating this issue as it pertains to population-based breast screening programs. Further studies are needed to determine if higher costs associated with the investment in machines, connectivity, data storage, and longer times for screen readings of DBT might be balanced against lower recall rates, less extensive diagnostic workup and treatment due to better or earlier detection.

Assessment of evidence table summary

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

Outcomes	Participants Studies	Quality of evidence	Overall results
Reduction in mortality	0	Nil	No studies reported on a reduction in mortality.
Cancer detection rate	1,293,328 participants Eight studies including two systematic reviews, one RCT and one fully paired study	⊕⊕⊕⊕ Strong	Pooled analysis of data from 17 studies plus new results from one RCT and final results from the OTS reported that FFDM+DBT increases cancer detection. The pooled incremental increase was 1.6 cancers per 1000 screening examinations. A higher CDR was reported for studies using double reading and biennial screening (2.4 cancers per 1000 screening examinations). Data from retrospective analysis confirm the direction of effect but some non-significant results were also presented in retrospective analyses, but this could be due to study effects.
Invasive cancer detection rate	151,776 participants Five studies including one RCT, one fully paired study	⊕⊕⊕⊕ Strong	Based on previous systematic reviews with pooled analysis and noting initial results from the RCT and final results from the OTS trial, use of FFDM+DBT resulted in statistically significant increases in detection of invasive cancers for women aged over 50 years. Inconsistent results were presented for DCIS detection.
Interval cancer detection	67,329 participants Three studies including two fully paired studies	⊕ Very low	No systematic review or pooled analysis was available. Primary studies reported non-significant results reported. No pooled analysis has been completed but meta-analysis is underway.
PPV_{1-3}	267,443 participants Seven studies including one RCT and two fully paired studies	⊕⊕⊕ Moderate	No systematic review or pooled analysis was available. Both primary studies observed statistically significant increases when DBT was used. Data from retrospective analysis confirm the direction of effect. No results were presented for PPV_2.
Recall rate	1,276,867 participants Eight studies	⊕⊕ Low	Pooled analysis suggests that recall rates decreases with the use of DBT, but the primary studies report inconsistent evidence.
False positive rate	45,408 participants Two studies	⊕ Very low	No systematic review or pooled analysis was available and primary study results are inconsistent.
Radiation dose	14,379 participants Three studies including one RCT and two fully paired studies	⊕⊕ Low	No systematic review or pooled analysis was available. Both primary studies observed an increase in MGD which was almost double that recorded for FFDM.
Interpretation time	19,560 participants One RCT	⊕ Very low	No systematic review or pooled analysis was available.

Table 2: Assessment of evidence for DBT+s2DM from all studies

Outcomes	Participants Studies	Quality of evidence	Overall results
Reduction in mortality	0	Nil	No studies reported on a reduction in mortality.
Cancer detection rate	327,114 participants Eight studies including one RCT	⊕⊕ Low	No systematic review or pooled analysis was available. Data from the RCT showed no statistically significant increase in CDR but all of the other paired studies or those set in European screening programs reported significant increases in CDR. Some of the American retrospective analyses did not report increases in CDR.
Invasive cancer detection rate	248,835 participants Eight studies including one RCT	⊕ Very low	No systematic review or pooled analysis was available. Data from the RCT showed no statistically significant increase in invasive CDR but all of the other studies reported increases but reported increases from two studies were descriptive only.
Interval cancer detection	53,268 participants One study	⊕ Very low	Non-significant results reported. No pooled analysis but meta-analysis underway.
PPV_{1-3}	294,347 participants Six studies including one RCT	⊕⊕⊕ Moderate	No systematic review or pooled analysis was available but all studies (including the RCT reported statistically significant increases in PPV_1 and PPV_3 (where this was reported). No results from prospective studies were presented for PPV_2.
Recall rate	616,334 participants Eight studies including one RCT	⊕⊕ Low	Pooled analysis suggests that recall rate decreases with the use of DBT, but the primary studies report inconsistent evidence.
False positive rate	69,745 Four studies	⊕ Very low	No systematic review or pooled analysis was available and primary study results are inconsistent.
Radiation dose	71,391 Five studies including one RCT	⊕ Very low	No systematic review or pooled analysis was available. Mixed results reported for radiation dose. No difference in MGD was reported in the RCT but data from all other studies indicated a higher MGD with DBT+s2DM (although the MGD for this modality was lower than for FFDM+DBT).
Interpretation time	86,755 participants Four studies	⊕⊕⊕ Moderate	No systematic review or pooled analysis was available but evidence that interpretation time increased with the use of DBT was consistent across all studies, with most reporting at least a doubling in reading time.

Table 3: Assessment of evidence for DBT_{MLO} compared to FFDM

Outcomes	Participants Studies	Quality of evidence	Overall results
Reduction in mortality	0	Nil	No studies reported on a reduction in mortality.
Cancer detection rate	14,848 participants One study	⊕ Very low	No systematic review or pooled analysis was available.
Invasive cancer detection rate	14,848 participants One study	⊕ Very low	No systematic review or pooled analysis was available.
Interval cancer detection	14,848 participants One study	⊕ Very low	Non-significant results reported. No pooled analysis but meta-analysis underway.
PPV_{1-3}	14,848 participants One study	⊕ Very low	No systematic review or pooled analysis was available.
Recall rate	14,848 participants One study	⊕ Very low	No systematic review or pooled analysis was available.
False positive rate	14,848 participants One study	⊕ Very low	No systematic review or pooled analysis was available.
Radiation dose	16,056 participants Two studies	⊕ Very low	No systematic review or pooled analysis was available. Mixed results were reported, with one study reporting a higher MGD with DBT_{MLO} and one reporting a lower MGD compared to FFDM.
Interpretation time	0	Nil	No studies reported on interpretation time.

1. INTRODUCTION

1.1. About digital breast tomosynthesis

Digital breast tomosynthesis (DBT) (also known as breast tomosynthesis, mammographic tomosynthesis or pseudo-three dimensional/3D mammography) is an imaging technology that can be used to detect, assess and diagnose breast cancer. DBT records low-dose images of a compressed breast. These images are reconstructed in (usually) 1mm parallel slices or 10mm slabs to form a three-dimensional image of the breast. Radiologists (or other readers) then analyse these images to determine the presence of 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 presenting as masses and increase lesion conspicuity for presentations like architectural distortion (AD). This potentially increases screening sensitivity and specificity as abnormalities are easier to see.

Radiation dose varies depending on the units used, whether the dose is automated or set by a radiographer, and by whether DBT is used alone, with integrated synthesized mammography (s2DM) image acquisition[1] or is used as an adjunct to full field digital mammography (FFDM[2]). 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 the evidence base underpinning the use of DBT+s2DM within a screening program grows.

1.2. BreastScreen Australia's position statement on tomosynthesis

The BSA program uses bilateral FFDM as the screening test for the early detection of breast cancer in asymptomatic women aged over 40 years. DBT is used in the examination of screen-detected abnormalities at some BSA assessment clinics. In 2014, the Community Care and Population Health Principal Committee of the Australian Health Ministers' Advisory Council (AHMAC) endorsed BSA's position statement on DBT. It concluded that FFDM remained the most effective population screening test 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. *Allen + Clarke* has previously undertaken narrative literature reviews on the emerging evidence about the role of DBT in screening and the assessment of screen-detected lesions (see *section 1.3*). These reviews have been considered by the Breast-Screening Technical Reference Group, which recommended that changes be made to the position statement. The key changes focus on supporting the use of DBT in the investigation of screen-detected abnormalities, but it does not recommend that DBT be used as the screening test at this time. A new position

[1] 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.
[2] FFDM is also known as two-view digital mammography.

statement on DBT was endorsed by the AHMAC Standing Committee on Screening in November 2019 and by the AHMAC Clinical Principal Committee in 2020.

1.3. Previous literature reviews undertaken by *Allen + Clarke*

Recognising that the evidence base underpinning the use of digital breast tomosynthesis DBT in breast cancer screening and the assessment and diagnosis of breast cancer is developing quickly, the Department of Health (Australia) commissioned two narrative literature reviews exploring the role of tomosynthesis in breast cancer screening of asymptomatic women. Key findings from both reviews are summarised in the report for ease of reference. Full texts are available at http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/dbt.

1.3.1. The role of DBT in as a screening test

In 2017, the Department of Health (Australia) contracted *Allen + Clarke* to undertake a literature review (not a 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. 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 screening test). This review considered if available published evidence on efficacy, effectiveness and safety indicated 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 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.

1.3.2. The role of DBT in the assessment clinic

Allen + Clarke undertook a second literature review on the role of DBT in the investigation of screen-detected abnormalities (i.e., its potential role in the assessment and diagnosis of breast cancer within a breast screening program). We wanted to know if DBT is a more sensitive and specific imaging technology that can contribute to reduced investigation for benign final outcome compared to other mammographic views including repeat FFDM, digital spot-compression or magnification views, or adjunctive ultrasound and adjunctive MRI. This review considered evidence assessing whether DBT should be the preferred method of assessment for women with a suspected malignancy identified by screening, and any incremental costs or safety considerations associated with the implementation of DBT as an imaging tool in assessment or diagnosis.

1.4. Purpose and scope of this literature review

Since the preparation of *Allen + Clarke*'s 2018 literature reviews, further evidence has been published (including final results from the Maroondah and Malmö trials and further results from the Oslo Tomosynthesis in Screening and STORM trials). Much of the newer evidence adds depth to what is already known, rather than shedding light on some of the unknown issues associated

with the use of DBT as a primary screening test. It is important that the updated BSA position statement on the role of DBT accurately reflects the literature base at the time of its publication even as further studies investigating longer-term screening outcomes continue. The Department of Health has engaged *Allen + Clarke* to update its 2018 literature reviews with research published to December 2019. This new narrative literature review will ensure that any updated BSA position statement on DBT most accurately reflects the current evidence base.

1.5. Ongoing research

Since *Allen + Clarke*'s 2018 literature review, several large trials have reported baseline, interim or final results including results from the Reggio Emilia and To-Be-1 RCTs and evaluation pilots/trials from key European and Australian screening programs (i.e., Malmö, Maroondah, Oslo, STORM-2, Trento and Verona); however, this is still an area of very active research. Our review of www.clinicaltrials.gov (completed on 22 November 2019) identified six large, ongoing studies investigating the role of DBT in population-based screening for asymptomatic women (either comparing FFDM to FFDM+DBT, DBT alone or DBT in combination with s2DM). Many of these studies are focusing on cost-effectiveness, pathologic and biological characteristics of screen-detected cancers (including cancer subtypes) and/or presentation of advanced cancers at a subsequent screening round, interval cancer rates and/or measures of mortality from breast cancer. Further updates to the BSA position statement may be required as these studies report interim or final findings. Key large, ongoing or upcoming trials include the:

- Tomosynthesis trial in Bergen (the To-Be-1 study in Norway, active status[3]; and a second RCT (the To-Be-2 study, recruiting status)[4]
- ProteusDonna trial (Piedmont, Italy, active status)[5]
- PROSPECTS trial (United Kingdom, recruiting status)[6]
- TOSYMA (Germany, recruiting status)[7], and
- Tomosynthesis Mammographic Imaging Screening trial (TMIST) (United States, recruiting status)[8].

[3] This study began in January 2016 and primary study completion is set for January 2020. Interim results are reported in this review.
[4] This is a five-year extension of the To-Be-1 trial and will cover one more screening round plus three years of follow-up. In this prospective cohort study, 32,400 women (est) will be screened with DBT+s2DM compared to prior screening with DBT+s2DM or FFDM alone (i.e., participants in To-Be-1). Key outcome measures focus on cancer detection rate, recall rates, PPV, prognostic and predictive tumour characteristics interval cancer rates, missed cancers, interpretation time, economic evaluation and pain/discomfort.
[5] This is a randomized controlled trial (RCT) of 92000 Italian women. Key outcomes include advanced cancers detected at the subsequent round, pathologic and biological characteristics of screen-detected cancers, interval cancers, cancer detection rate, recall rate, false positive rate, PPV, and cost-effectiveness analysis.
[6] This is a prospective randomized controlled trial of 100,000 women aged 50-70 years recruited from National Health Service Breast-screening Program sites. The trial will investigate the impact and cost effectiveness of FFDM+DBT (including s2DM images) compared to FFDM alone (the current screening test). A non-inferiority test will be applied to FFDM+DBT compared to DBT+s2DM. Key outcomes include sensitivity, prognostic features, cancer size at detection, interval cancer rate and modelling of cancer mortality reduction. It will also look at impact by population group. The study is due for completion in July 2024. Michael Michell is the lead investigator.
[7] This prospective, randomized trial of 80,000 asymptomatic women began recruiting in March 2018. It will investigate DBT+s2DM compared to FFDM. Study completion is proposed for July 2023. Key outcome measures focus on cancer detection rates including by cancer type and category, interval cancer rate, recall rate and PPV.
[8] This is a randomized phase III trial of 164,946 women. Initial results are expected in 2025 with final results reported in 2030. Women will undergo annual screening with either DBT or FFDM. Outcome measures include women diagnosed with advanced breast cancer, BIRADS imaging features, pathologic and biological characteristics of screen-detected cancers (including cancer sub-type), pathologist agreement, breast cancer mortality, image quality, diagnostic and predictive performance, PPV, NPV, sensitivity and specificity, interval cancer, recall rates, biopsy rates, and healthcare cost.

Additionally, there may be further papers published from data collected in the Maroondah pilot, which could explore issues relating to DBT and overdiagnosis in more detail and within the context of the BSA program.

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 builds on two previous literature reviews on DBT conducted by *Allen + Clarke.* The same research questions and methodology were used, enabling effective comparison of results from new research with previous findings.
- This is not a systematic review. We have provided statements about the quality of the evidence included in this review, but no primary research or pooled analysis was undertaken.
- The following database were searched on 11-13 November 2019: OVID Medline. The following websites were reviewed: clinicaltrials.gov, the Cochrane database, National Institute for Health and Clinical Excellence, UK National Institute for Health Research HTA database, and the UK NHSBPS.
- All returned citations and abstracts were assessed for relevance to the research questions, PICO(T/S) criteria and inclusion in previous *Allen + Clarke* literature reviews. 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. This replicates the methodology from the previous literature reviews.
- A total of 41 articles met the inclusion criteria.
    - Two systematic reviews covering immediate screening outcomes
    - Five narrative literature reviews or editorial commentary by experienced researchers
    - Four primary research papers from two RCTs
    - Nine prospective studies embedded in a European population-based screening program
    - Two prospective cohort studies embedded in a European population-based screening program that used a historical cohort
    - Three sub-studies from the STORM-2 trial
    - One prospective study embedded in an American screening program
    - Three observer performance studies
    - 11 retrospective analyses, and
    - One cost-effectiveness study.

2.1. Objectives

We wanted to understand how the evidence on DBT's role as a primary screening test has developed since 1 January 2018 (the closing date for *Allen + Clarke*'s previous literature reviews[9]). It aims to determine whether FFDM remains the best test for the early detection of clinically significant 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 cancer early. That is, confirming whether digital mammography is still the best test for the early detection of breast cancer and what (if any further) updates are needed to BSA's position statement on DBT as a screening tool. Specific topics include:

- the efficacy, effectiveness and safety of DBT as a screening tool (which will support the consideration of whether DBT should be the preferred method for screening asymptomatic women for breast cancer in Australia)
- whether DBT should be the preferred method for screening asymptomatic women with dense breasts or women aged between 40 and 50 years, and
- the incremental costs associated with implementing DBT as a screening tool and patient and health practitioner experience with this screening modality as well as any other specific evidence about how to implement DBT into the BSA screening program as the primary test.

A systematic review with pooled analysis was not performed.

2.2. Research questions

2.2.1. Questions about effectiveness, efficacy and safety

The three questions about effectiveness, efficacy 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(T/S) criteria underpinning these research questions are described in *Table 6* (overleaf).

[9] http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/dbt

Table 6: PICO(T/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 risk factors for breast cancer including familial history/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 of breast lesions) Interval cancer rates Surrogate mortality indicators (tumour size at detection, lymph node negativity, grade)
Study types	Systematic reviews, RCT

2.2.2. Question about implementation

The question 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(T/S) criteria underpinning these research questions are described in *Table 7* (below).

Table 7: PICO(T/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

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(T/S) criteria underpinning these research questions are described in *Table 8* (below).

Table 8: PICO(T/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

2.3. Literature search

The following databases were searched on 11-13 November 2019:

- Cochrane Library database
- National Institute for Health and Clinical Excellence
- OVID Medline
- UK National Institute for Health Research HTA database, and
- UK NHSBPS.

The website www.clinicaltrials.gov was searched on 22 November 2019.

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 (i.e., 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 2018 – 31 December 2019), English language, human 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).

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., its potential role in intra-operative imaging)
- compared DBT to screening strategies other than FFDM
- focused on a study population other than asymptomatic women (i.e., diagnostic studies that focused on cancer detection in symptomatic women or other diagnostic imaging)
- related to refinement in technique, image processing, or CAD (not linked to reading time)
- focused on measuring breast density or used phantoms, or
- related to post-treatment surveillance.

We excluded 14 studies published between 1 January 2018 and 31 May 2018 because these had previously been covered comprehensively in *Allen + Clarke*'s 2018 literature review on the role of tomosynthesis in the assessment clinic.[10] We included one study covered in this previous

[10] Citations from these studies are listed here: **Ariaratnam** NS, Little ST, Whitley MA, Ferguson K. (2018). Digital breast Tomosynthesis vacuum assisted biopsy for Tomosynthesis-detected Sonographically occult lesions. *Clinical Imaging*, *47*, 4–8. https://doi.org/10.1016/j.clinimag.2017.08.002; **Bahl** M, Gaffney S, McCarthy AM, Lowry KP, Dang PA, Lehman CD. (2018). Breast Cancer Characteristics Associated with 2D Digital Mammography versus Digital Breast Tomosynthesis for Screening-detected and Interval Cancers. *Radiology*, *287*(1), 49–57. https://doi.org/10.1148/radiol.2017171148; **Bahrs** SD, Otto V, Hattermann V, Klumpp B, Hahn M, Nikolaou K, Siegmann-Luz K. (2018). Breast tomosynthesis for the clarification of mammographic BIRADS 3 lesions can decrease follow-up examinations and enables immediate cancer diagnosis. *Acta Radiologica (Stockholm, Sweden: 1987)*, *59*(10), 1176–1183. https://doi.org/10.1177/0284185118756458; **Caumo** F, Romanucci G, Hunter K, Zorzi M, Brunelli S, Macaskill P, Houssami N. (2018). Comparison of breast cancers detected in the Verona screening program following transition to digital breast tomosynthesis screening with cancers detected at digital mammography screening. *Breast Cancer Research And Treatment*, *170*(2), 391–397. https://doi.org/10.1007/s10549-018-4756-4; **Destounis** S. (2018). Role of Digital Breast Tomosynthesis in Screening and Diagnostic Breast Imaging. *Seminars In Ultrasound, CT, And MR*, *39*(1), 35–44. https://doi.org/10.1053/j.sult.2017.08.002; **Dibble** EH, Lourenco AP, Baird GL, Ward RC, Maynard AS, Mainiero MB. (2018). Comparison of digital mammography and digital breast tomosynthesis in the detection of architectural distortion. *European Radiology*, *28*(1), 3–10. https://doi.org/10.1007/s00330-017-4968-8; **Endo** T, Morita T, Oiwa M, Suda N, Sato Y, Ichihara S, Arai, T. (2018). Diagnostic performance of digital breast tomosynthesis and full-field digital mammography with new reconstruction and new processing for dose reduction. *Breast Cancer (Tokyo, Japan)*, *25*(2), 159–166. https://doi.org/10.1007/s12282-017-0805-9; **Förnvik** D, Kataoka M, Iima M, Ohashi A, Kanao S, Toi M, Togashi K. (2018). The role of breast tomosynthesis in a predominantly dense breast population at a tertiary breast centre: Breast density assessment and diagnostic performance in comparison with MRI. *European Radiology*, *28*(8), 3194–3203. https://doi.org/10.1007/s00330-017-5297-7; **Garayoa** J, Chevalier M, Castillo M, Mahillo-Fernández I, Amallal El Ouahabi N, Estrada C, Valverde J. (2018). Diagnostic value of the stand-alone synthetic image in digital breast tomosynthesis examinations. *European Radiology*, *28*(2), 565–572. https://doi.org/10.1007/s00330-017-4991-9; **Heywang-Köbrunner** SH, Hacker A, Jänsch A, Kates R, Wulz-Horber S, German Reader Team. (2018). Use of single-view digital breast tomosynthesis (DBT) and ultrasound vs. Additional views and ultrasound for the assessment of screen-detected abnormalities: German multi-reader study. *Acta Radiologica (Stockholm, Sweden: 1987)*, *59*(7), 782–788. https://doi.org/10.1177/0284185117732600; **Houssami** N. (2018). Evidence on Synthesized Two-dimensional Mammography Versus Digital Mammography When Using Tomosynthesis (Three-dimensional Mammography) for Population Breast Cancer Screening. *Clinical Breast Cancer*, *18*(4), 255-260.e1. https://doi.org/10.1016/j.clbc.2017.09.012; **Michell** MJ, Batohi B. (2018). Role of tomosynthesis in breast imaging going forward. *Clinical Radiology*, *73*(4), 358–371. https://doi.org/10.1016/j.crad.2018.01.001; **Pan** HB, Wong KF, Yao A, Hsu GC, Chou CP, Liang HL, Yang TL. (2018). Breast cancer screening with digital breast tomosynthesis—4 year experience and comparison with national data. *Journal Of The Chinese Medical Association: JCMA*, *81*(1), 70–80. https://doi.org/10.1016/j.jcma.2017.05.013; **Rodriguez-Ruiz** A, Gubern-Merida A, Imhof-Tas M, Lardenoije S, Wanders AJT, Andersson I, Sechopoulos I. (2018). One-view digital breast tomosynthesis as a stand-alone modality for breast cancer detection: Do we need more? *European Radiology*, *28*(5), 1938–1948. https://doi.org/10.1007/s00330-017-5167-3;

literature review (Phi et al., 2018) because it included some material about the role of DBT as a screening test which had not previously been captured.

2.3.1. Validation

To determine if the search retrieved the correct range of available research, a validation process was completed using three recent literature reviews/overview commentaries by leading researchers relevant to the primary research questions: Chong et al., 2019, Lång, 2019, and Rocha García & Fernández, 2019.

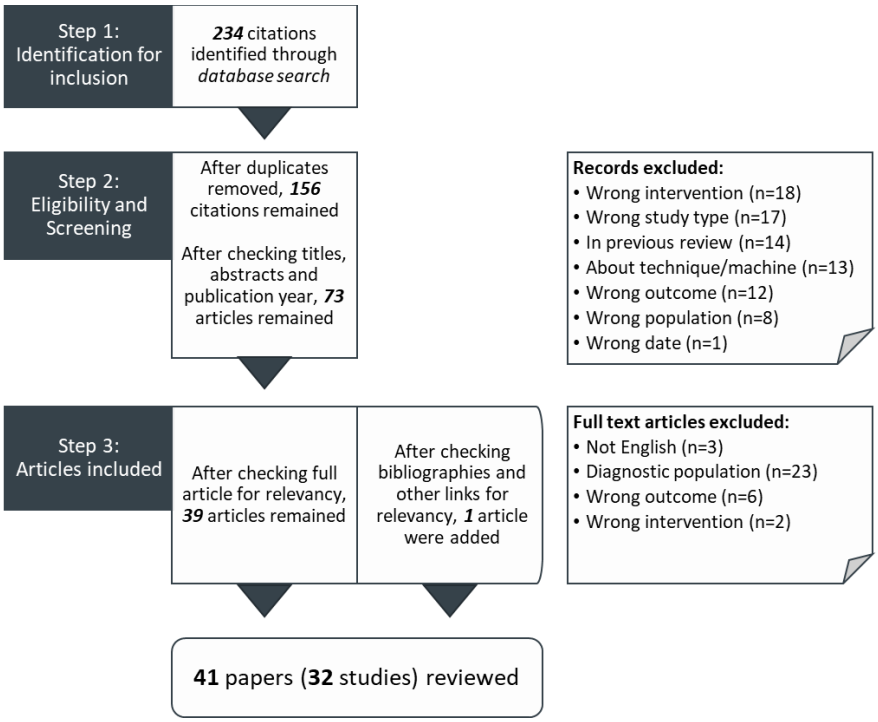
There was a high degree of consistency between in the studies returned using our strategies and those included in the three reviews; however, only a small number of studies included within these papers were published after 1 January 2018. To counter this, we reviewed www.clinicaltrials.gov to check whether papers from recently completed trials had been published and, if so, whether our search had captured these. We also completed bibliography checks of each included paper and are confident that this review is comprehensive for studies that have a high degree of applicability to the BSA program.

2.3.2. Inclusion process

From a first sweep based on citation and abstract, full texts for all proposed inclusions were retrieved and reviewed for relevance to the research questions, inclusion criteria and documented PICO(T/S) 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



In order to ensure greatest use of this literature review for the BSA program, we deliberately prioritised reporting on studies that provided the most relevant data to the Australian screening context (i.e., studies set in population-based screening with a biennial screening interval and double reading). We have, of course, included a number of studies that have different policy settings (eg, annual screening interval and/or single reading), but have not reported on these to the same extent. The list of papers by study type is listed below.

- Two systematic reviews covering immediate screening outcomes (Marinovich et al., 2018; Phi et al., 2018)
- Five narrative literature reviews or editorial commentary by experienced researchers (Chong et al., 2019; Lång, 2019; Rocha García & Fernández, 2019; Zackrisson, 2019; Li et al., 2018)
- Four papers from two RCTs (the To-Be-1 RCT and baseline data from the Reggio Emilia RCT) (Aase et al., 2019; Hofvind et al., 2019; Moger et al., 2019; Pattacini et al., 2018)
- Nine prospective studies embedded in a European population-based screening programs (Houssami et al., 2019; Johnson et al., 2019; Miglioretti et al., 2019; Skaane et al., 2019; Hofvind et al., 2018; Østerås et al., 2018; Romero Martín et al., 2018; Skaane et al., 2018; Zackrisson et al., 2018)
- Two prospective cohort studies embedded in a European population-based screening program that used a historical cohort (Bernardi et al., 2019; Caumo et al., 2018)
- Three sub-studies from the STORM-2 trial (Bernardi et al., 2018; Gennaro et al., 2018; Houssami et al., 2018)
- One prospective study embedded in an American screening program (Rose & Shisler, 2018)
- Three observer performance studies (Chae et al., 2019; Choi et al., 2019; Lai et al., 2018)
- 11 retrospective analyses (Ambinder et al., 2019; Bahl et al., 2019; Conant et al., 2019; Dang et al., 2019; Fuiji et al., 2019; Honig et al., 2019; Hovda et al., 2019; Simon et al., 2019; Wasan et al., 2019; Upadahay et al., 2018; Wahab et al., 2018), and
- One cost-effectiveness study (Lowry et al., 2019).

A further nine studies from *Allen + Clarke*'s previous literature reviews are also referenced, including the pooled analysis prepared by Phi et al. (2018), Coop et al. (2016), and Hodgson et al. (2016), three papers from the STORM trial and a technical evaluation of DBT (Strudley et al., 2014).

2.3.3. Imaging systems used in the literature

We assessed the imaging systems and screening strategies used in all studies included in this literature review to understand whether the findings are broadly generalizable or whether findings relate to specific units/machinery.

Imaging units used in studies reported in this literature review varied, reflecting the dynamic nature of breast screening technology at this time. Hologic units were the most commonly used DBT-capable units.

A list of DBT-capable units and the studies they were used in is summarised below:

- Hologic Selenia Dimensions units (20 papers: Bahl et al., 2019; Bernardi et al., 2019; Choi et al., 2019; Conant et al., 2019; Fuiji et al., 2019; Houssami et al., 2019; Hovda et al., 2019; Simon et al., 2019; Skaane et al., 2019; Wasan et al., 2019; Bernardi et al., 2018; Caumo et al., 2018; Gennaro et al., 2018; Hofvind et al., 2018; Houssami et al., 2018; Østerås et al., 2018; Romero Martín et al., 2018; Rose & Shisler et al., 2018; Skaane et al., 2018; Wahab et al., 2018)
- GE Senographe Essential, SenoClair or Pristina units (seven papers including papers from both RCTs: Aase et al., 2019; Chae et al., 2019; Hovda et al., 2019; Hofvind et al., 2019; Moger et al., 2019; Hofvind et al., 2018; Pattacini et al., 2018)
- Siemens Mammomat Inspirations (four papers: Hovda et al., 2019; Johnson et al., 2019; Hofvind et al., 2018; Zackrisson et al., 2018), and
- Not stated but covered multiple sites so likely to involve multiple units (two papers: Honig et al., 2019; Miglioretti et al., 2019).

This literature update reviewed the imaging software used to create the synthesised images. Many studies did not include details of the software used but all that did used Hologic's C-View (i.e., the evaluation pilots in Cordoba, Maroondah and the STORM 2 trial and a study by Simon et al., 2019).

2.3.4. Imaging protocols used in the primary studies

Studies reported in this literature review used a wide range of different imaging protocols. Compared to *Allen + Clarke*'s earlier literature review on DBT in screening, more recent research observed performance of DBT+s2DM compared to FFDM alone (in the literature published to December 2017, research focused on more on dual acquisition as DBT+s2DM was only emerging as technique of interest). Since then, screening programs that have chosen to implement DBT (eg, Trento and Verona) and the only Australia pilot (Maroondah) have used DBT+s2DM rather than FFDM+DBT. Primary imaging techniques were:

- FFDM compared to DBT+s2DM: 12 papers
- FFDM compared to FFDM+DBT: 12 papers
- FFDM compared to DBT_{MLO}: two papers, and
- Other combinations: six papers.

The studies by Hovda et al. (2019) and Romero Martín et al. (2018) reported on complex sequences of imaging, grouping of study participants and/or reading strategies. These studies were generally designed to test whether DBT alone (either one view or single reading) had good short-term performance metrics or could be used to decrease reading time. We have chosen to primarily report results from these studies based on the arms that are most relevant to the way that programs are implementing DBT into screening programs (i.e., DBT+s2DM compared to FFDM alone or two sequential FFDM screening examinations compared to FFDM followed by DBT+s2DM or FFDM+DBT). Imaging protocols are summarised in *Table 5* (overleaf). Large bolded (✓) refer to a study's main imaging protocol (noting that some studies or sub-studies reported some data on additional reading protocols).

Table 5: Imaging protocols used in the primary studies included in this literature review

Key papers	Study	FFDM alone	DBT+ s2DM	FFDM+ DBT	DBT_{MLO}	s2DM alone	DBT alone	Other protocols
Aase at al. (2019); Hofvind et al. (2019); Moger et al. (2019)	To-Be-1	✓	✓					
Bernardi et al. (2018); Gennaro et al. (2018)	STORM-2	✓	✓	✓		✓		Single projections
Bernardi at al. (2019)	Trento	✓	✓					
Caumo et al. (2018)	Verona	✓	✓					
Houssami et al. (2019)	Maroondah	✓	✓					
Romero Martín et al. (2018)	Cordoba	✓	✓					FFDM+s2DM +DBT
Hofvind et al. (2018)	Norway BSP	✓	✓					
Hovda et al. (2019)	Norway BSP	✓	✓	✓				
Lai et al. (2018)		✓	✓					
Pattacini et al. (2018)	Reggio Emilia	✓		✓				
Houssami et al. (2018)	STORM	✓		✓				
Skaane et al. (2019); Østerås et al. (2018); Skaane et al. (2018)	OTS	✓	✓	✓			✓	CAD
Conant et al. (2018)	PROSPR	✓		✓				
Rose & Shisler (2018)	31 US sites	✓		✓				
Bahl et al. (2019)		✓		✓				
Dang et al. (2019)		✓		✓				
Fuiji et al. (2018)		✓		✓				
Honig et al. (2019)	4 US sites	✓		✓				
Wasan et al. (2018)		✓		✓				
Ambinder et al. (2018)			✓	✓				
Simon et al. (2019)			✓	✓				
Johnson et al. (2019); Zackrisson et al. (2018)	Malmö	✓			✓			
Miglioretti at al. (2019)		✓					✓	
Chae et al. (2019)							✓	DBT+CAD
Choi et al. (2019)		✓	✓	✓		✓		
Wahab et al. (2018)		✓				✓		

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 (biennially screening for invited women aged 50-74 years, with double reading by trained radiologists). Specifically, we are interested in understanding if DBT (either alone or integrated with other mammography) 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. The results are presented by the following clinical outcomes and performance metrics:

- Sensitivity: overall cancer detection rate (CDR), interval cancer rate and relative sensitivity including the impact DBT may have on CDR for prevalent and incident screening or for sub-populations of women (i.e., women with heterogeneously dense or extremely dense breasts or by age)
- Cancer type and histopathological characteristics
- Radiological presentation
- Specificity: overall recall rate, false positive recall rate, relative specificity and PPV), and
- Safety (i.e., radiation dose).

Most research focused on shorter-term performance measures like CDR, recall rate and cancer type rather than longer-term screening outcomes such as mortality reduction or a reduction in interval cancers. Uncertainty about the effectiveness of DBT as the primary test in a population-based screening program remains (and is likely to remain until larger long-term studies report results). That said, DBT+s2DM has been implemented as the primary screening test in several European and American screening programs, with initial results reporting improvement in short-term outcomes for program sensitivity.

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 overall CDR, interval cancer rate and relative sensitivity. Cancer type and tumour characteristics at detection are also very important to understand whether clinically significant cancers are being detected (see *section 3.2*).

We want to know, based on current evidence, what role DBT plays on a modern breast-screening environment and which screening strategy (DBT+s2DM, FFDM+DBT or other imaging) is best able to detect the most clinically relevant invasive breast cancers and at an early stage/low grade. *Section 3.1* describes overall CDR, interval cancer detection and relative sensitivity.

3.1.1. Overall cancer detection rate

Overall CDR refers to the total number of cancers that can be identified using a specific imaging technique(s). It is a fundamental short-term outcome and marker of a screening program's effectiveness at detecting very small cancers that are not currently presenting symptomatically.

Information about whether DBT detects more invasive disease compared to DCIS (including information on invasive CDR) is discussed in *section 3.2*.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

Most studies compared FFDM to FFDM+DBT (i.e., adjunctive DBT). Results from studies reporting on DBT+s2DM compared to FFDM or FFDM+DBT were emerging and other imaging combinations were being trialled to reduce radiation dose (i.e., DT_{MLO}+DM_{CC}).

There was strong evidence that CDR increased when using FFDM+DBT compared to FFDM alone. All studies with adequate statistical power to detect small changes to CDR reported that use of DBT increased cancer detection. Increases were reported in a range of studies (including large prospective trials with a paired design and set in European population-based screening programs). Only one small early retrospective analysis did not report an increase in overall CDR. While different results were achieved for different combinations of screening strategy, the direction of effect is consistent across study design, setting and location. There was variance in effect estimates. Further, reported results for DBT+s2DM included CDR that were at least as good as FFDM+DBT but this approach was delivered with a reduced radiation dose. DBT+s2DM was identified as a very promising approach.

CDR data from prospective trials

Results from large, fully paired prospective trials embedded in population screening programs consistently reported statistically significant incremental CDR of more than two per 1000 biennial screening examinations. Pooled analysis based on data from the Screening with Tomosynthesis or Regular Mammography (STORM) and Oslo Tomosynthesis in Screening (OTS) trials reported a statistically significant increase of 2.43 cancers detected per 1000 screening examinations compared to FFDM alone. The incremental CDR and invasive CDR were similar for FFDM+DBT and DBT+s2DM: either approach detected significantly more cancers than FFDM alone. The Malmö trial reported a significant increase in CDR: 2.6 cancers detected per 1000 screening examinations using DBT_{MLO}+DM_{CC} compared to FFDM alone.

DBT+s2DM also performed better, detecting 8.8 cancers per 1000 screening examinations compared to 6.3 with FFDM alone. 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).

CDR data from retrospective analysis or reader studies

Data from retrospective studies showed a similar effect (that is, CDR increased when DBT was 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 most likely related to differences in reading strategy (eg, double reading compared to a single reader approach or to annual rather than biennial screening), participant selection, under-powering or other study design limitations.

Most of the studies investigating CDR had short time frames ($\leq$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 any changes in interval cancer rate) and if the improvement in CDR provided by DBT (either with integrated s2DM or with FFDM) is sustained between first and subsequent (i.e., prevalent and incident) screening examinations.

The previous GRADE assessment (below) reflected the strong evidence that FFDM+DBT increased CDR but that the evidence was still emerging for other imaging combinations.

GRADE assessment: cancer detection rate: FFDM+DBT compared to FFDM alone

Participants Studies	Quality of evidence	Overall results
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).

GRADE assessment: cancer detection rate: DBT+s2DM compared to FFDM+DBT or FFDM alone

Participants Studies	Quality of evidence	Overall results
153,668 participants 5 studies	⊕⊕ Low	No systematic review or pooled analysis was available.

GRADE assessment: cancer detection rate: DBT_{MLO} compared to FFDM alone

Participants Studies	Quality of evidence	Overall results
7,681 2 studies	⊕ Very low	No systematic review or pooled analysis was available.

Updated findings

Literature published since 31 December 2017 reported on overall CDR. It included new pooled analysis (completed on studies published to July 2017), results from the first RCT of DBT in a 'live' screening setting (Hofvind et al., 2019), final data from some of the large European trials including the OTS and Malmö trials and results from pilot evaluations in Europe and Australia. There was a mix of single-site and multicentre studies. This section provides a brief update of study findings on overall CDR. Most of the studies reporting on CDR also described the histological and pathological characteristics of cancers detected (including whether detected cancers were invasive or not). This information further develops our understanding of whether the additional cancers detected with DBT are clinically significant or whether they are small, slow-growing invasive cancers or ductal carcinoma in situ (DCIS) for which prognosis is good even if detected at a later screening round (i.e., overdiagnosis). Information about the types of cancer detected, nodal status, grade and size is provided in *section 3.2*. Information about CDR stratified by age, prevalent/incident screening round and breast density is provided in *sections 1.3.2* to *1.3.4*.

Primary studies already incorporated into the systematic reviews (Marinovich et al., 2018; Phi et al., 2018) were generally published prior to 2018 and were reviewed in *Allen + Clarke*'s previous literature reviews on the role of DBT in screening and the assessment of screen-detected lesions. Relevant data from other primary studies published since then is included in evidence tables (see *Table 9a, 9b* and *9c*). Included studies are listed below.

Systematic review and literature review

Two systematic reviews with pooled analysis: Marinovich et al., 2018; Phi et al., 2018

Two narrative literature reviews: Chong et al., 2019; Li et al., 2018

One editorial: Zackrisson, 2019

Randomized controlled trials

One RCT (one paper): Hofvind et al., 2019

One randomized study embedded in a population screening program: Pattacini et al., 2018

Prospective studies embedded in population screening programs

Eight studies: Bernardi et al., 2019; Houssami et al., 2019; Skaane et al., 2019; Caumo et al., 2018; Hofvind et al., 2018; Romero Martín et al., 2018; Skaane et al., 2018; Zackrisson et al., 2018

Observational studies

Seven studies with a retrospective design: Bahl et al., 2019; Conant et al., 2019; Dang et al., 2019; Fuiji et al., 2019; Hovda et al., 2019; Ambinder et al., 2018; Rose & Shisler 2018

Summary of key findings published between 1 January 2018 and 31 December 2019

The findings add further depth to *Allen + Clarke*'s 2018 literature reviews but some mixed results were reported in the To-Be-1 RCT (the first large RCT of DBT). Pooled evidence from papers published before December 2017 (involving more than one million women) reported a statistically significant increase in CDR, with an incremental increase in detection of 1.6 cancers per 1000 screening examinations. Higher increases were reported in European population-based programs with similar policy settings to the BreastScreen Australia program (i.e., pooled analysis indicated 2.4 more cancers were detected when DBT was used).

Studies published since 1 January 2018 are generally consistent with the pooled analysis (that use of DBT significantly increases CDR and that DBT+s2DM is not inferior to FFDM+DBT or FFDM alone), with one important exception. The To-Be-1 RCT found no statistically significant increase in CDR with the use of DBT+s2DM. Possible explanations for this difference were that all previous images were available to readers, which may have favoured FFDM imaging (and which probably better reflects a 'real world' screening environment). There may be other study differences to which require further unpicking. That said, evidence is still strong that use of DBT increases cancer detection.

Increases in CDR were reported in all other studies (including large, ongoing robust prospective trials and robust retrospective analyses) for different combinations of screening strategy including FFDM+DBT, DBT+s2DM, and DBT_{MLO} compared to FFDM (the Malmö trial). Evidence

also suggests that DBT+s2DM is superior to FFDM alone for CDR and is not inferior to FFDM+DBT. For this reason, where screening programs have chosen to implement DBT, they have chosen DBT+s2DM rather than dual acquisition. Other imaging protocols were also explored (including DBT_{MLO} or single reading of DBT), again favouring the use of DBT in terms of increasing cancer detection.

Evidence is also emerging that DBT increases CDR in women with more dense breasts and older women. Generally, evidence suggested that use of DBT results in increased CDR for all breast densities (i.e., all women) but some evidence is emerging that the increase may be greater for women whose breasts are more dense. While breast density is not reported in the BSA program, considering results of stratification analysis by density is useful in considering whether there are some population groups for whom DBT may be more beneficial.

Literature review

Chong et al. (2019) completed a narrative literature review on DBT, reporting results on CDR. Again, the studies reported in Chong et al. reflect those discussed in *Allen + Clarke*'s 2018 literature review on the role of DBT in screening as well as three papers published since on the Verona pilot (Caumo et al., 2018), the BreastScreen Norway program (Hofvind et al., 2018) (both DBT+s2DM), and the Reggio Emilia pilot (Pattacini et al., 2018) (FFDM+DBT). Chong et al. reported results consistent with the meta-analysis undertaken by Marinovich et al. (2018) and Phi et al. (2018): FFDM+DBT increases CDR compared to FFDM alone, with a range of 1.2 to 4.6 additional cancers detected reported. For the prospective European trials (paired design), CDR increased more (incremental CDR per 1000 screening examinations ranged from 1.9 to 4.1) than for the retrospective American studies (incremental CDR per 1000 screening examinations ranged -0.8 to 1.9). More information about the Verona, Reggio Emilia and Norway studies is provided in the following section.

Systematic reviews

We identified two systematic reviews with pooled analysis of overall CDR (Marinovich et al., 2018; Phi et al., 2018) and two narrative literature reviews (Chong et al., 2019; Li et al., 2018). Results from the meta-analysis reported consistent increases in CDR with the use of DBT as part of the screening test. Reported incremental CDR per 1000 screening examinations ranged between -0.8 fewer cancers detected per 1000 screening examinations (a non-significant result from an early study and the only study to report a lower CDR when DBT was used in the imaging protocol) to 3.9 cancers detected per 1000 screening examinations (a statistically significant result from the Verona pilot, p=.001).

Pooled analysis (Marinovich et al., 2018; Phi et al., 2018) drew only on studies published prior to 31 December 2017. All the studies reported in these papers were reported in *Allen + Clarke*'s previous literature reviews. The Marinovich et al. and Phi et al. papers reported data from the paired trials in different ways. Data from the Malmö, STORM and STORM-2 trials[11] was reported separately from the retrospective American studies (which compared FFDM to FFDM+DBT) but it is not clear in the method write-ups whether and how the different imaging protocols in the paired trials were accounted for in the pooled analysis (specific notes are provided below).

[11] The Malmö trial compared one view DBT (DBT_{MLO}) to FFDM. The STORM trial compared FFDM to FFDM+DBT. The STORM-2 trial compared FFDM to FFDM+DBT and DBT+s2DM.

Because of this, we report the results of each pooled analysis here rather than under the imaging protocol.

Marinovich et al. (2018) completed a meta-analysis of 17 studies comparing the role of DBT in cancer detection in asymptomatic women attending for breast-screening. Results were presented by a grouping based on paired or unpaired study design. Paired studies included the four main prospective European trials embedded in population-based screening programs published between 2009 and July 2017 (i.e., interim results from the Malmö trial and data from the OTS, STORM and STORM-2 trials). Unpaired studies included the main retrospective American studies. Different imaging protocols were included in the Malmö and STORM-2 trials[12], but all other studies compared FFDM+DBT to FFDM alone. Different reading protocols and screening intervals were also used. The total number of study participants was 1,009,790 (37,092 women in paired studies; 972,698 women in unpaired studies). Included studies reported on CDR and recall rate/false recall rate and were designed to enable comparison between FFDM and FFDM+DBT. To complete the meta-analysis, Marinovich et al. used QUADAS-2 to undertake the quality assessment and reported a low risk of bias in the included studies; however, the included studies were also quite different in some fundamental ways (screening interval, screening intervention/comparator, etc.). No analysis by prevalent or incident screening was undertaken as this information was not available from the primary studies. No studies involving high-risk or symptomatic women or tomosynthesis in other settings were included.

While results from Marinovich et al.'s meta-analysis draw on data older than the inclusion criteria for the current literature review, it provided a helpful summary of the evidence as it stood shortly before *Allen + Clarke* completed its first literature review on the role of DBT in screening (rather than describing advances in knowledge since 31 December 2017). For CDR, Marinovich et al.'s paper reached the same conclusion that *Allen + Clarke* did: FFDM+DBT detects more cancers than FFDM alone. Results were consistent between paired and unpaired studies although the effect estimates differed (see box, right). The pooled (paired/unpaired results) was an incremental increase in CDR of 1.6 cancers per 1000 screening exams. Differences between paired/unpaired data may reflect underpinning differences in the screening process (eg, single reading and an annual screening interval is used in most of the retrospective analysis, but the paired trials used biennial screening and double reading). Marinovich et al., noting these differences, proposed that DBT use may have a greater impact in biennial screening given the longer screening interval.

Study Design	**Results: overall CDR** (95%CI)
Marinovich et al., 2018 Meta-analysis including data from Malmö, OTS and STORM	CDR for paired studies: • FFDM+DBT: 8.8 (7.4, 10.5) • FFDM alone: 6.4 (5.2, 7.9) • Pooled increase for FFDM+DBT: 2.4 (1.9, 2.9) CDR for unpaired studies: • FFDM+DBT: 5.7 • FFDM alone: 4.5 • Pooled increase for FFDM+DBT: 1.1 (0.8, 1.5) Pooled incremental increase • FFDM+DBT: 1.6 (1.1, 2.0)
Phi et al., 2018 Systematic review with meta-analysis including data from Malmö and STORM	CDR for paired studies: • RR=1.52 (1.08, 2.12) CDR for unpaired studies: • RR=1.33 (1.2, 1.47)

[12] Although the CDR data presented in Marinovich et al.'s 2018 paper is the data for FFDM and FFDM+DBT imaging. The s2DM results are not included.

Phi et al. (2018) completed a systematic review with meta-analysis to understand the relationship between DBT compared to FFDM for women with dense breasts. The authors identified 16 studies published between May 2007 and May 2017, 11 of which were set in a primary screening environment and reported on CDR. Most of these studies were completed in the United States (n=6) but papers from the STORM and Malmö trials were also included. All these papers were included in *Allen + Clarke*'s 2018 review on DBT in screening and are subject to the same differences in study design, intervention/comparator and screening interval as described for Marinovich et al.'s study. Also, like Marinovich et al., Phi et al. reported that pooled CDR estimates from all studies was higher when DBT was compared to FFDM alone, but that CDR improved more in pooled data from the paired trials (Malmö and STORM data):

- For American studies (annual screening): RR=1.33 (95%CI: 1.20, 1.47).
- For paired European trials (biennial screening): RR=1.52 (95%CI: 1.08, 2.12).

No pooled CDR estimates by breast density were reported in Phi et al.'s paper.

Pooled analysis demonstrated consistent evidence that DBT increases cancer detection, with a larger increase in CDR reported where the screening interval is longer and where double reading is used (i.e., a greater increase in CDR is seen in double-read, biennial screening compared to annual screening with single reading). This is relevant to the BSA program, which uses both double reading and offers screening biennially.

DBT+s2DM compared to FFDM

Having been identified as a promising approach, most of the studies published since 31 December 2017 compare DBT+s2DM to FFDM alone, reflecting practices designed to reduce the radiation dose associated with dual acquisition if FFDM+DBT is used. Studies included in *Allen + Clarke*'s previous literature reviews and the pooled analysis published since then (Marinovich et al., 2018; Phi et al., 2018) featured studies comparing FFDM to FFDM+DBT. Therefore, this new imaging protocol represents a shift to a different way of integrating DBT into a screening program.

RCT

While there have been a number of fully paired studies set in population-based screening programs, the To-Be-1 trial is the first RCT in a screening program (the Norway BreastScreen program). Three papers have been published on this RCT on the use of DBT+s2DM as a screening test (a main analysis by Hofvind et al., 2019; two sub-studies by Aase et al., 2019 and Moger et al., 2019).

The To-Be-1 RCT aimed to assess the screening performance of DBT+s2DM. Study participants were women born between 1947 and 1966 who attended for screening. About 25 percent of possible participants did not attend and none of the papers commented on how screening non-attendees differed from attendees. Study participants were randomized to receive either DBT+s2DM (14,380 women) or FFDM alone (14,369 women) using purpose-built software. Women with breast implants, breast symptoms or a previous history of breast cancer or metastatic melanoma were excluded. Randomization occurred after consent to participate in the study was received but the intervention was not blinded. Only one screening test was performed during the RCT and each mammogram was double-read (independent, with consensus) with prior results for all mammograms completed in the prior 10 years (including of any diagnostic

investigation) available to all readers. Reading order was standardized (s2DM image then DBT 10mm slabs then DBT 1mm slices in CC and MLO views).

A total of 182 cancers were detected. The To-Be-1 RCT reported different results from previous paired studies, almost all of which reported incremental increases in CDR (two to three additional cancers per 1000 screening examinations).

Hofvind et al. (2019) found no statistically significant difference in overall CDR between the two study arms:

> *CDR for DBT+s2DM: 6.6 cancers detected per 1000 women screened*
>
> *CDR for FFDM: 6.1 per 1000 women screened (RR=1.09; 95%CI: 0.82, 1.46, p=.56).*

Information about type of cancer detected in this RCT are discussed in *section 3.2*.

The results for the FFDM imaging are consistent with other results reported for studies set in population-based screening programs; however, the CDR results for DBT+s2DM are much lower than what has been reported in the double-read, biennial programs. Given that the results of this RCT were inconsistent with almost all previous study results on CDR (including the paired studies), Hofvind et al. (2019) and Zackrisson (2019) commented on possible reasons for this result. Hofvind et al. noted that a lower radiation dose used (mean glandular dose of 2.96mGy) may have affected adversely image quality on the DBT+s2DM imaging (although this is a higher MGD compared to the MGD reported in other studies using DBT+s2DM). Other possible reasons for the difference in CDR (and other) results proposed by Hofvind et al. (2019) included:

- reader training (DBT+s2DM was a novel technique for some readers at the start of this RCT but no information on sensitivity by reader was provided and a learning effect was noted through an increase in consensus between readers over time with DBT+s2DM: presumably in other studies readers would have had a similar learning curve, see *section 4.2*)
- the availability of all prior mammograms in the To-Be-1 study (which may not have been available in all previous studies or which are not available for women attending for a prevalent screen), which may have influenced readers' decisions on FFDM
- the potential influence of the reading protocol: reading of s2DM images first or reviewing DBT slabs before slices may have decreased sensitivity if potentially suspicious areas were not identified during slabbing and readers looked through the slices more quickly (Zackrisson, 2019), or
- another underlying study effect such as machinery used (i.e., the machines were first generation GE units) or the underpowering of the study (Hofvind et al. noted that a multi-centre sample of 400,000 women was needed in order to detect statistically significant differences).

No information was provided about differences in CDR by prevalent/incident screening round (and therefore the availability of prior mammogram images). Given the study was underpowered, the results require careful interpretation. We conclude that DBT+s2DM was as safe as FFDM but Hofvind et al.'s results suggest that it may not be a superior screening test. Further large RCTs are needed to confirm whether the CDR results presented in the To-Be-1 RCT can be replicated.

Prospective studies embedded in population-based screening programs

Since December 2017, the following pilot evaluations and trials comparing DBT+s2DM to FFDM have been completed:

- Maroondah (Australia) (Houssami et al., 2019)
- OTS (Norway) (Skaane et al., 2019)
- Trento (Italy) (Bernardi et al., 2019)
- Cordoba (Spain) (Romero Martín et al., 2018)
- A three-centre prospective cohort study in the Norway BreastScreen Program (Hofvind et al., 2018), and
- Verona (Italy) (Caumo et al., 2018).

Results from these prospective studies embedded in screening programs consistently demonstrated increased CDR for the DBT imaging protocol, which aligns with earlier findings but are not consistent with To-Be-1 RCT data. Effect estimates varied from 8.67 to 9.8 cancers detected per 1000 screening examinations with DBT+s2DM compared to 5.41 to 6.6 cancers detected per 1000 screening examinations with FFDM. Below are short summaries of the key study methodologies and findings from CDR. Each of these studies also described cancer type (invasive/DCIS) and characteristics, which are discussed in *section 3.2*.

The Maroondah pilot trial (Houssami et al., 2019) provided the first Australia-specific evidence regarding the use of DBT+s2DM in the BSA program. This pilot study, undertaken in the Maroondah BreastScreen service, included 10,146 women attending for screening: 4993 women had DBT+s2DM; 5153 women received screening with FFDM. Study participants were assigned to be screened by DBT+s2DM or FFDM based on machine availability and booking schedule. A small number of study participants (n=38) attended for annual screening, all other women attended for biennial screening. More women attending for their first breast screen were screened with DBT+s2DM (19.5%) compared to FFDM alone (6.8%). Women screened with DBT+s2DM were also younger (58.0 years compared to 62.3 years). All images were double-read in line with the BSA National Accreditation Standards, with arbitration by a third reader in cases of disagreement between the first two readers. In the Maroondah trial, 49 cancers were detected. DBT+s2DM detected more cancers than FFDM:

- DBT+s2DM: 9.8 cancers per 1000 screening examinations (95%CI: 7.2, 13)
- FFDM: 6.6 cancers per 1000 screening examinations (95%CI: 4.6, 9.2), and
- Estimated difference in CDR: 3.2 more cancers with DBT+s2DM (95%CI:-.32, 6.8).

No statistical testing was completed (i.e., it is an estimate only).

Both the Trento and Verona Screening Programs were STORM trial sites. Following the completion of STORM, both screening programs adopted DBT as the screening test and evaluated its implementation in pilot evaluations by comparing results to historical cohorts (Bernardi et al., 2019; Caumo et al., 2018). In the Trento study, women aged over 50 years participating in screening from 15 October 2014 and 14 October 2016 were screened with DBT+s2DM. The results were compared with a historical cohort of women enrolled in the program in the years immediately preceding the pilot. Women who had participated in STORM were excluded, resulting

in 46,343 women receiving DBT+s2DM compared to a historical cohort of 37,436 women screened with FFDM. Following implementation, more cancers were detected with DBT+s2DM compared to FFDM alone (8.67 per 1000 screening examinations compared to 5.48 per 1000 screening examinations).

Like the Trento evaluation study, the Verona prospective study investigated DBT+s2DM in a screening program, comparing results to a historical cohort of women from the same program (Caumo et al., 2018). Women who are high risk (eg, with a BRCA mutation) or a personal history of breast cancer do not participate in this program. Overall CDR for DBT+s2DM was 9.3 cancers per 1000 screening examinations compared to 5.41 cancers per 1000 screening examinations for FFDM (RR1.72; 95%CI: 1.30, 2.29). These results are similar to the Trento and Maroondah studies.

Two prospective studies embedded in population-based screening programs have also assessed whether DBT+s2DM is feasible as a screening test (compared to FFDM). In their prospective transversal reading study set in the Cordoba Screening Program, Romero Martín et al. (2018) evaluated different reading strategies to determine whether single reading of DBT+s2DM was feasible as a screening test (i.e., if FFDM was not required at all). Four reading arms were completed (including two single FFDM read and one double FFDM read, single reading of DBT+s2DM and FFDM+s2DM+DBT). We report only the results from the double reading of FFDM and DBT+s2DM arms. Romero Martín et al. reported study results for 16,068 screens. CDR improved with a single-read DBT+s2DM compared to double-read FFDM, with an incremental cancer detection of 12.6 percent favouring DBT+s2DM. This incremental improvement was lower than in other studies discussed in this section. This finding may relate to the single reading of the DBT+s2DM arm compared to double reading of FFDM but it also indicated that single reading of DBT+s2DM may not be inferior to double-read FFDM. In addition, there was no statistically significant increase in CDR between DBT+s2DM+FFDM, indicating that DBT+s2DM may not require FFDM imaging in order to detect the same number of cancers (thus reducing overall reading time – see *section 4.3* for more results).

Other studies also reported consistent results with these trials/evaluations, including a prospective cohort study set in three Norwegian screening programs (Oslo, Vestfold and Vestre Viken), which reported a CDR of 9.4 cancers per 1000 screening examinations when DT+s2DM was used compared to 6.1 cancers per 1000 FFDM screening examinations (Hofvind et al., 2018).

The OTS trial (described in the following subsection on FFDM+DBT compared to FFDM alone) reported final results in 2019 (Skaane et al., 2019). Skaane et al. observed that DBT+s2DM had similar short-term screening outcomes compared to FFDM+DBT (and that it is therefore not inferior). While this paper did not report on CDR, it reported on sensitivity: 69 percent (95%CI: 64.3, 76.2) for DBT+s2DM compared to 70.5 percent (95%CI: 63.1, 75.1, *p*=.77) for FFDM+DBT (see *section 3.1.6* for further discussion of this point).

From the more robust data presented in this section, it appears that DBT+s2DM is not inferior to FFDM alone as a screening test and that s2DM may perform as well as FFDM in the imaging protocol in terms of cancer detection. In addition, CDR is used as a measure of radiologist performance, with most radiologists improving CDR when using DBT+s2DM imaging.

Retrospective analysis

A small retrospective analysis of American data (7845 FFDM+DBT mammograms compared to 14,776 DBT+s2DM mammograms) explored whether DBT+s2DM was inferior to FFDM+DBT,

reporting that there was no statistically significant difference between the two imaging protocols (Ambinder et al., 2018).

FFDM+DBT compared to FFDM alone

Fewer studies reported results comparing FFDM+DBT to FFDM alone compared to the number published in the last literature review, reflecting the change in imaging protocol used to reduce radiation dose (i.e., the research is shifting to DBT+s2DM to reflect practice). In addition, the only robust studies discussed in this section are the presentation of final results from the OTS trial and baseline data for half of the participants in the Reggio Emilia RCT.

RCT

Pattacini et al. (2018) reported a baseline interim analysis of the 19,560 women participating in the Reggio Emilia DBT RCT, an Italian two arm, test and treat trial. In this RCT, women are randomized to either the FFDM+DBT (9777 women) or FFDM arm (9783 women) for the first screening examination followed by a future biennial screening examination with FFDM only. Follow-up is for 4.5 years. Women cannot participate in the RCT if they have not participated in screening before, if they have a history of breast cancer, have previously had DBT, have very large breasts, have breast implants, are pregnant or are high risk for breast cancer. The recruitment rate for this study suggests that it is underpowered to detect interval cancers (one of the primary study outcomes). The interim analysis included the CDR for FFDM+DBT (8.6 cancers per 1000 screening examinations) compared to 4.5 per 1000 screening examinations for FFDM alone (RR=1.89; 95%CI: 1.31, 2.72). Further specific information about this study is provided in *Table 9b*.

Prospective studies embedded in population-based screening programs

Reporting on the OTS trial (a fully paired study embedded in a population-based screening program involving results for 24,301 women), Skaane et al. (2018) reported final results for the FFDM+DBT arm compared to FFDM alone. Skaane et al. reported statistically significant increased CDR with the use of FFDM+DBT, with incremental cancer detection of 3.0 cancers per 1000 screening examinations (95%CI: 1.7, 4.4, p<.001). While the overall CDR was a little higher than other studies (9.3 cancers per 1000 screening examinations), increases are similar to those reported in the studies published before December 2017 as well as some of the results achieved with DBT+s2DM.

Retrospective analysis

Hovda et al. (2019) completed a complex retrospective analysis of paired data from the Norway BreastScreen Program, comparing results from sequential rounds of screening with either FFDM followed by DBT or two rounds of DBT. Hovda et al. separated paired data into four groups based on the screening tests different groups of women had:

- Group 1 women had two sequential FFDM screening examinations (n=10,502)
- Group 2 women had a first FFDM screening examinations followed by FFDM+DBT or DBT+s2DM (n=29,262)
- Group 3 women had FFDM+DBT then were screened with DBT+s2DM (n=8799), and
- Group 4 women had FFDM+DBT then FFDM or FFDM then DBT+s2DM (n=20,631).

To ensure consistency with the BSA program, we report only on the results comparing Group 1 and Group 2. Hovda et al. reported a statistically significant increase in CDR when DBT was used (9.9 cancers per 1000 screening examinations were detected in the DBT Group 2 arm compared to 4.6 per 1000 screening examinations when two consecutive FFDM examinations were performed, *p*=.001).

Conant et al. (2019) completed a large, multi-site cohort study (50,971 DBT examinations; 129,369 FFDM examinations on women aged 40-74 years) within the PROSPR consortium to determine if short-term screening outcomes (like CDR and recall rates) varied by imaging, age and breast density. Overall, data from the three centres indicated that DBT imaging resulted in higher CDR (OR 1.41; 95%CI: 1.05, 1.89, *p*=.02). These differences were observed across all age groups and breast densities. Specific differences in CDR reported by age group and breast density are discussed in further detail in *section 3.1.3*.

We identified one American observational study (Fuiji et al., 2019) which reported no increase in CDR when DBT was used (compared to FFDM); however, the paper noted that several different imaging protocols were used during the study period (i.e., FFDM alone, FFDM+DBT at the start of the study then DBT+s2DM or FFDM+DBT+s2DM). Results analysis in the paper did not separate the data by imaging protocol and the authors did not identify this as a limitation, noting that the lack of an incremental increase in CDR with DBT could be due to a CDR of 5.0 cancers per 1000 screening examinations for FFDM, which is high for an annual, single reader program.

Three other retrospective analyses (Bahl et al., 2019; Dang et al., 2019; and Rose & Shisler, 2018) also reported on CDR. Dang et al. (2019) performed a matched retrospective cohort study comparing CDR for FFDM+DBT to FFDM alone. Most participants were aged under 50 years, and the cohort included 23,997 screening examinations (14,180 were FFDM, 9817 were FFDM+DBT). The focus of this study was to measure changes in CDR and by tumour type. The authors reported statistically significant increases in CDR when FFDM+DBT was used, increasing CDR by 1.9 per 1000 screening examinations (*p*=.01) (although the reported increases were not as large as reported in the European matched cohort studies):

- FFDM+DBT: 3.7 per 1000 screening examinations
- FFDM alone 1.8 per 1000 screening examinations

Bahl et al. (2019) completed a review of consecutive mammograms (n=71,958) for women aged over 65 years prior to the integration of DBT into an American mammography clinic. They reported no statistically significant differences in CDR between the two imaging modalities for this sub-group of women, reporting that DBT was not inferior to FFDM. More information on this study is provided in *section 1.3.4*. Rose & Shisler's analysis also focused on age and investigated the impact of DBT on CDR in women aged under 50 years. As seen in some of the data presented on DBT+s2DM, there is no consistent evidence describing this impact as the increase in CDR was not significant (increasing from 1.9 per 1000 screening examinations to 2.6 with the use of DBT, *p*=.06).

DBT_{MLO} compared to FFDM

The Malmö trial used a different imaging protocol to reduce radiation dose, compression and reading time associated with FFDM+DBT (Zackrisson et al., 2018). Instead of FFDM+DBT or DBT+s2DM, 14,484 women aged 40-76 years participating in the Malmö trial were screened with

FFDM then with wide-angle one-view DBT (DBT_{MLO}). Previous mammograms were available to readers. Reading steps were:

- Current FFDM images then previous FFDM images if available then recording breast density (FFDM reading group), or
- DBT_{MLO} then DM_{CC} then previous FFDM images if available (DBT reading group).

A total of 21,691 women were invited to participate in the Malmö trial (68 percent uptake rate). The authors did not provide any information about the women who chose not to participate (noting only that they did not come to screening or chose to have FFDM only), making it difficult to determine the full applicability of trial results to the general population. All images were double-read (as per the Swedish Breast-screening Programme's protocols, with a five-point rating used). Information about breast density was recorded using BIRADS categories. The follow-up period was either 18 months or 24 months depending on the woman's age (with younger women followed-up at the earlier time). Interim results from this trial were presented in *Allen + Clarke*'s first literature review on screening. Final results (including CDR and interval cancer rates) are reported in this current literature review.

A total of 139 cancers were detected:

- 89 cancers were detected with FFDM and DBT_{MLO}
- 42 cancers were only detected with DBT_{MLO}, and
- Eight cancers were only detected with FFDM (but were identified on DBT imaging although not recalled for further assessment – see *section 3.3* for further discussion on the radiologic presentation of these cancers).

The DBT reading arm had a higher CDR: 8.7 cancers detected per 1000 screening examinations compared to 6.5 cancers detected for the FFDM reading arm (p=.0001). This resulted in a statistically significant incremental CDR of 2.2 cancers per 1000 screening examinations. Although the screening strategy and DBT system used in this trial differs from other studies, final results from the Malmö trial confirm the direction of effect for CDR seen in other studies and suggest that there is evidence that different imaging protocols can deliver non-inferior performance compared to FFDM (although non-inferiority testing was not completed for this study).

Table 9a: DBT+s2DM compared to FFDM alone: studies reporting on overall CDR and invasive CDR

Study	Sample	Study type	**DBT+ s2DM** 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 alone** Invasive CDR per 1000 screening examinations (95%CI; p=value)	**Incremental detection** per 1000 screening examinations (95%CI)
RCT							
Hofvind et al. (2019) To-Be-1 (main analysis)	28,749 women aged 50-69y participating in BreastScreen Norway (Bergen) DBT+s2DM: 14,380 FFDM alone: 14,369	Single site RCT of one screening round, imaged on GE SenoClair with independent double reading (n=8 radiologists)	6.6 (5.3, 7.9)	6.1 (4.4, 7.3)	*Invasive* 5.6 (4.3, 6.8) *DCIS* 1.0 (0.5, 1.6)	*Invasive* 4.9 (3.8, 6.1; *p*=.47) *DCIS* 1.1 (0.6, 1.7; *p*=.86)	No data
Prospective studies embedded in population-based screening programs							
Bernardi et al. (2019) Trento	DBT+s2DM: 46,343 participants in the Trento (Italy) screening program (mean age = 57.9y) FFDM alone: historical cohort (n=37,436) previously screened in the Trento program (mean age = 57.9y)	Prospective single site pilot evaluation following the implementation of DBT into the Trento screening program imaged with Hologic systems with independent double reading (n=8 radiologists)	8.67	5.48	*Invasive (#)* 352 cancers *DCIS (#)* 50 DCIS (12.47% of cancers detected)	*Invasive (#)* 173 cancers *DCIS (#)* 33 DCIS (16.10% of cancers detected)	RR=1.58 (1.34, 1.87)
Houssami et al. 2019) Maroondah	10,146 Australian women: • DBT+s2DM: 4993 women (5018 screening examinations); mean age 58.0y • FFDM alone: 5153 women (5166 screening examinations); mean age 62.3y	Prospective single site pilot trial embedded in a screening program (Maroondah) imaged on Hologic Selenia Dimensions 8000 (+/- C-view software) or Siemens Mammomat Inspiration with independent double reading (n=7 radiologists)	9.8 (7.2, 13)	6.6 (4.6, 9.2)	No CDR data but 40 invasive cancers and nine in-situ cancers detected	No CDR data but 30 invasive cancers and four in-situ cancers detected	3.2 (-0.32, 6.8) more cancers with DBT+s2DM
Caumo et al. (2018) Verona	DBT+s2DM: 16,666 participants in the Verona screening program; median age = 59y	Prospective single site pilot evaluation following the implementation of DBT into the Trento screening program imaged	9.30	5.41	*Invasive (#)* 136 cancers *DCIS (#)* 19 DCIS	*Invasive (#)* 58 cancers *DCIS (#)* 20 DCIS	*Overall* RR=1.72 (1.30, 2.29)

Study	Sample	Study type	DBT+ s2DM 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 alone Invasive CDR per 1000 screening examinations (95%CI; p=value)	Incremental detection per 1000 screening examinations (95%CI)
	FFDM: historical cohort of 14,423 women previously screened in the Verona program; median age = 58y	with Hologic systems with independent double reading (n=4 radiologists)			(12.3% of cancers detected)	(25.6% of cancers detected, *p*=.002)	
Hofvind et al. (2018) Norway BSP	DBT+s2DM: 37,185 attendees at an Oslo clinic (7250 had a prevalent screen), mean age=59.2y FFDM: 61,742 attendees at the Vestfold or Vestre Viken clinics (9517 had a prevalent screen), mean age=59.4y	Prospective, multi-site, population-based cohort study imaged on Hologic Dimensions, Siemens Mammomat Inspirations or GE SenoEssential units with independent double reading (n=24 radiologists)	9.4	6.1 (*p*<.001)	Total cancers: 348 *Invasive* 7.6 *DCIS* 1.7	Total cancers: 379 *Invasive* 5.3 (*p*<.001) *DCIS* 0.8 (*p*<.001)	No data but authors observed statistically significant increases (see other columns)
Romero Martín et al. (2018) Cordoba	16,067 women in the Cordoba Screening Program (Spain) (aged 57.59y) undergoing FFDM+DBT with s2DM images (3341 women attending for prevalent screening; 127,27 for incident screening)	Prospective transversal reading study comparing single and double reading of FFDM, DBT+s2DM (single read) and FFDM+DBT+s2DM (single read), imaged on Hologic Dimensions unit; n=5 radiologists	5.4 (69 cancers detected; 18 only detected with this reading)	4.7 (69 cancers detected; seven only detected with this reading)	71 invasive cancers (no CDR prepared) 21 in-situ cancers (no CDR prepared)	57 invasive cancers (no CDR prepared, (*p*=.001) 19 in-situ cancers (no CDR prepared) (*p*=.774)	12.6% (7.2, 21.2; *p*=.043)
Retrospective analysis							
Hovda et al. (2019) Norway BSP	35,736 women with at least two consecutive screens in the Oslo BreastScreen Norway program Groups included FFDM then FFDM (n=10,502) compared to FFDM then FFDM+DBT (n=8365) plus FFDM then DBT+s2DM (n=13,059) (total in DBT arm = 21,424)	Retrospective paired analysis of screening round data (first round was negative at screening or assessment; second round could be negative or positive) using Hologic Dimensions and GE Senographe systems with independent double reading	9.9	4.6 (*p*<.001)	*Invasive* 8.4 *DCIS* 1.6	*Invasive* 3.7 *DCIS* 0.9 (*p*=.001)	No data

Table 9b: FFDM+DBT compared to FFDM alone: studies reporting on overall and invasive 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; p=value)	Other data (95%CI)
RCT							
Pattacini et al. (2018) Reggio Emilia	19,560 women aged 45-74y attending for incident screening (women aged 45-49y screened annually; women aged ≥50 screened biennially). Women with family history were excluded. FFDM: 9783 women (mean age=56.3y) FFDM+DBT: 9777 women (mean age=56.2y)	Baseline data from a prospective three-centre RCT using a test and treat methodology using GE Senographe Essential systems	All: 8.6 Women aged ≤49y: 5.2 Women aged 50-59y: 9.0 Women aged 60-70y: 10.5	All: 4.5 Women aged ≤49y: 2.9 Women aged 50-59y: 4.4 Women aged 60-70y: 5.8	*Invasive* 7.1 *DCIS* 14.0	*Invasive* 4.0 (RR=1.77; 95%CI: 1.20, 2.62) *DCIS* 5 (RR=2.80; 95%CI: 1.01, 7.65	*Relative risk and risk difference* All: 1.89 (1.3, 2.72); 4 (2, 6) Women aged ≤49y: 1.83; 2 (-2, 6) Women aged 50-59y: 2.04 (1.18, 3.58); 5 (1, 8) Women aged 60-70y: 1.83 (1.05, 3.20); 5 (0-9)
Prospective studies set embedded in population-based screening programs							
Skaane et al. (2018) OTS	24,301 women aged presenting for screening in Oslo, Norway	Prospective, single site fully paired, double-reading trial using Hologic Dimensions system (FFDM, DBT and s2DM images) and GE Senographe (previous screening round FFDM images) and four reading arms	9.3	6.3	*Invasive* 84.1% of cancers *DCIS* 15.9% of cancers	*Invasive* 79.9% *DCIS* 20.1% (*p*=.234)	*Overall CDR (Incremental detection)* 3.0 (1.7, 4.4; *p*<.001)

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)	Other data (95%CI)
Retrospective analysis							
Bahl et al. (2019)	23,997 screening examinations in women aged 54y (±10.4y) FFDM: 14,180 FFDM+DBT: 9817	Retrospective matched cohort study (single site) using a range of units (not specified but units were replaced throughout the analysis)	3.7	1.8	75%	72% (*p*=.86) Invasive RR=2.2; 95%CI: 1.2, 3.9, *p*=.01	Overall CDR OR=2.1 (1.3, 3.5; *p*=.001) Incremental CDR: 1.9
Conant et al. (2019) PROSPR consortium	96,269 women aged 40-74y participating in annual screening through the PROSPR consortium (USA), mean age=55.7y FFDM: 129,369 screening examinations (mean age 56.4y) FFDM+DBT: 50,971 screening examinations (mean age 54.6y)	Retrospective, three site analysis with DBT images collected on Hologic Selenia Dimensions unit, varying DM units (not recorded) with single reading (n=2 radiologists)	*All women* 5.82 *Women aged 40-49y* Not dense: 4.41 Dense: 5.20 *Women aged 50-64y* Not dense: 4.03 Dense: 7.59 *Women aged 65-74y* Not dense: 9.64 Dense: 9.58	*All women* 4.42 *Women aged 40-49y* Not dense: 2.71 Dense: 2.93 *Women aged 50-64y* Not dense: 3.68 Dense: 5.51 *Women aged 65-74y* Not dense: 6.75 Dense: 7.63	No data	No data	OR=1.41 (1.05, 1.89, *p*=.02)
Dang et al. (2019)	23,997 screening examinations FFDM: 14,180 exams (26 cancers) FFDM+DBT: 9817 exams (37 cancers)	Retrospective matched cohort analysis of all breast screening exams at single site between October 2012 and September 2014, 24 readers	3.7	1.8	2.8	1.3	*Overall* RR=2.1 (1.3, 3.5, *p*=.01) *Invasive* RR=2.2 (1.2, 3.9, *p*=.01)
Fuiji et al. (2019)	Participants: 66,003 women DBT: 86,349* FFDM alone: 97,378 ** it is not clear which DBT imaging protocols were used*	Retrospective observational study set in the Vermont BreastScreen program using Hologic Selenia Dimensions units with 49 readers	5.0	5.6	No data	No data	Adjusted OR=0.94 (0.78, 1.14)

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)	**Other data** (95%CI)
Rose & Shisler (2018)	59,921 screening examinations of women aged under 50y (no lower age range provided, no stratification), estimate of 10 percent attending for prevalent screening examination: FFDM: 41,542 examinations FFDM+DBT: 18,379 examinations	Multi-site (N=31) retrospective study set in community mammography practices imaged on Hologic Selenia Dimensions units with independent double reading (n=7 radiologists)	*All women* 2.6 *Women with dense breasts (ACR heterogeneously and extremely dense)* 3.5	*All women* 1.9 *Women with dense breasts (ACR heterogeneously and extremely dense)* 2.1	*All cancers* 49 cancers *Invasive cancers* 33 cancers *In situ cancers* 16	*All cancers* 77 cancers *Invasive cancers* 50 cancers *In situ cancers* 27	*Incremental difference – all* 0.8 (-0.03, 1.6, *p*=.06) *Incremental difference density* 1.3 (0.1, 2.5, *p*=.03)

Table 9c: DBT_{MLO} compared to FFDM

Study	Sample	Study type	**DBT_{MLO}** CDR per 1000 screening examinations (95%CI; p-value)	**FFDM** CDR per 1000 screening examinations (95%CI; p-value)	**DBT_{MLO}** Number of cancers detected	**FFDM** Number of cancers detected	**Incremental detection** (95%CI)
Randomized controlled trials and trials embedded in population-based screening programs							
Zackrisson et al. (2018) Malmö trial	14,848 women aged 40-76y (mean age = 57y) presenting for screening in Malmö, Sweden *NB women aged 40-54 were screened every 18m; women aged 55-74y were screened every 24m.*	Prospective, single site paired trial where women underwent FFDM and wide-angled one-view DBT_{MLO} using Siemens Mammomat Inspirations	8.7 (7.3, 0.3)	6.5 (5.2, 7.9; *p*=.0001)	Invasive: 114 In-situ: 17 *Only detected by DBT_{MLO}:* Invasive: 38 In-situ: 4	Invasive: 80 In-situ: 17 *Only detected by DBT_{MLO}:* Invasive: 4 In-situ: 4	2.2 (2, 3.2; *p*<.0001)

3.1.2. CDR stratified by prevalent/incident screening round

Prevalent screening refers to the first time a woman participates in breast-screening. At this time, recall rates are often higher as benign conditions may be identified and require investigation. CDR can also be higher in a prevalent screening round as this is the first time a breast may have been imaged. Incident screening refers to all subsequent screening examinations. No systematic reviews of studies published since December 2017 considered CDR stratified by prevalent or incident screening round. Four of the trials/pilot evaluations stratified CDR by prevalent and incident screening round. Detailed methodologies for each of these studies is provided in *section 3.1.1.* Study results presented in these papers advance our knowledge as papers published before 31 December 2017 did not report on this.

DBT+s2DM compared to FFDM+DBT or FFDM alone

Prospective studies embedded in population-based screening programs

Four prospective studies reported on differences in CDR by prevalent or incident screening round, but no consistent findings were reported (i.e., some studies reported no significant differences between CDR for prevalent and incident rounds whereas other studies reported more cancers were detected with DBT at incident screening rounds compared to FFDM). Data is summarised below and methodologies for each of the following studies are described in *section 3.1.1*:

- Cordoba Screening Program: non-significant results for CDR were reported for women undergoing a prevalent screen (comparing DBT+s2DM to FFDM alone), but a statistically significant increase in CDR was reported for women undergoing an incident screening examination (an increase of 22.1 percent; 95%CI: 13.8, 33.2, *p*=.001) (Romero Martín et al., 2018)
- Maroondah pilot evaluation: most cancers were detected in the incident round for both DBT+s2DM and FFDM (40/49 for DBT+s2DM and 31/34 for FFDM), with similar CDR for prevalent and incident screens but a higher CDR for prevalent screening with DBT+s2DM was observed (Houssami et al., 2019):

	Prevalent screening examination *CDR per 1000 screening examinations (95%CI)*	**Incident screening examination** *CDR per 1000 screening examinations (95%CI)*
DBT+s2DM	9 (4, 17)	9.9 (7.1, 14)
FFDM	9 (2, 24)	6.4 (4.4, 9.1)

- Trento: DBT+s2DM increased CDR for both prevalent and incident screening but the authors did not compare prevalent and incident results to each other to see if the differences between the groups were significant and they did not report by age and prevalent screen:
  - Prevalent: 8.64 cancers per 1000 DBT+s2DM screening examinations compared to 5.93 per 1000 FFDM screening examinations; RR=1.46 (95%CI: 0.99, 2.15)

- Incident: 8.68 cancers per 1000 DBT+s2DM screening examinations compared to 5.38 per 1000 FFDM screening examinations; RR=1.61 (95%CI: 1.34, 1.94) (Bernardi et al., 2019).

- Verona: DBT+s2DM increased CDR for incident screening only (RR for subsequent screening = 1.77; 95%CI: 1.29, 2.44).

The results reported above suggested that that there may a greater benefit conferred to women in having incident screening with DBT+s2DM (rather than their first screen). There are a couple of possible reasons for this including that cancers may be more conspicuous in women undergoing a first screening round but that malignant changes may become more subtle in time and more common with increasing age (and women undergoing incident screening are likely to be older).

Another aspect of CDR noted by Marinovich et al. (2018) was that most of the studies published prior to July 2017 (their literature close date) focused on the first round of screening with DBT (and that these could be considered to be a 'prevalent' round). They noted the importance of further evidence on CDR to see whether the incremental increases seen with the use of DBT remained for following screening rounds using DBT. This is not what has been observed in the studies published since then including the Trento and Verona studies (which were also STORM sites) but further consistent evidence of this effect is required.

Retrospective analysis

We only identified one paper (Hovda et al., 2018) that described a retrospective analysis of Norway BreastScreen Program data. The authors compared short-term screening outcomes for several different scenarios including two consecutive screening rounds with FFDM and two or three consecutive screening rounds with FFDM followed by FFDM+DBT or DBT+s2DM. The study included records from 35,736 women who had at least two consecutive screening rounds between 2008 and 2016. Imaging protocols and associated CDR included:

- Group 1 (two consecutive FFDM screens): 4.6 cancers per 1000 screening examinations
- Group 2 (FFDM then FFDM+DBT or FFDM then DBT+s2DM): 9.9 cancers per 1000 screening examinations
- Group 3 (FFDM+DBT then DBT+s2DM): 8.3 cancers per 1000 screening examinations, and
- Group 4 (FFDM+DBT then FFDM or FFDM then DBT+s2DM): 4.3 cancers per 1000 screening examinations.

This study provided some initial data suggesting that CDR remains elevated with two consecutive rounds of screening with DBT compared to two rounds of screening with FFDM. Given the limited available information, the relationship between increased CDR and incident/prevalent screening (in this context) requires further investigation.

3.1.3. CDR stratified by breast density

One of the key advantages conferred by DBT is the use of very thin slices/slabs which enables radiologists to scroll through the breast images in a way that minimises any 'noise' associated with breast tissue overlap seen on two-dimensional mammography. Seven studies reported data that stratified CDR by breast density (usually comparing less dense breasts with more dense breasts). Detailed methodologies for each of these studies is provided in *section 3.1.1.*

Generally, evidence suggested that use of DBT results in increased CDR for all breast densities (i.e., all women) but some evidence is emerging that the increase may be greater for women whose breasts are more dense. While breast density is not reported in the BSA program (and was therefore not covered by the Maroondah trial), considering results of stratification analysis by density is useful in considering whether there are some population groups for whom DBT may be more beneficial.

Systematic reviews

Phi et al. (2018) completed a systematic review with pooled analysis reporting on breast density and CDR. Primary studies included in the pooled analysis generally used radiologist assessment to determine breast density (rather than using an automated system like Volpara). This could lead to some inconsistency in allocation of women to having scattered density or heterogeneously dense breast tissue. Phi et al. also included studies of screening and diagnostic populations, which could also influence the study's findings. That said, Phi et al. reported that CDR increased, and recall reduced for all breast density categories when DBT was used. Larger increases were reported for the more dense breast categories in prospective studies set in screening programs (OR=1.52; 95%CI: 1.08, 2.11). This indicated that DBT is likely to reduce 'noise' associated with breast tissue overlap and improve conspicuity of both benign and malignant lesions.

Marinovich et al. (2018) also looked at incremental CDR in their meta-analysis and noted the infrequency of data reported by breast density. They reported a non-significant increase in incremental CDR for women with dense breasts (BIRADS C+D) when DBT was used (based on DBT+FFDM imaging).

RCT

Pattacini et al. (2018) mentioned CDR and breast density in passing, noting that CDR increased for all categories of breast density with FFDM+DBT was used. Incremental increases in CDR were reported as follows, with only a significant result reported for scattered density:

- BIRADS A: RR=2.0 (95%CI: 0.37, 10.92) FFDM+DBT compared to FFDM (almost entirely fat)
- BIRADS B: RR=2.1 (95%CI: 1.05, 4.15) FFDM+DBT compared to FFDM (scattered density)
- BIRADS C: RR=1.5 (95%CI: 0.83, 2.72) FFDM+ DBT compared to FFDM (heterogeneously dense breasts)
- BIRADS D: RR=2.29 (95%CI: 0.94, 5.56) FFDM+DBT compared to FFDM (extremely dense).

Prospective studies embedded in population-based screening programs

CDR was also stratified by breast density in two trials (Cordoba and Verona) (Caumo et al., 2018; Romero Martín et al., 2018). While breast density was classified using BIRADS (5th edition) in these studies, data was stratified differently (either less dense compared to more dense or by each density category). The effect trend is similar regardless of stratification approach: use of DBT+s2DM increased CDR in all densities but the increase was greater in more dense breasts. In the Cordoba trial, Romero Martín et al. (2018) stratified data into four separate categories (A,B,C,D). Most cancers were detected in breasts with scattered or heterogenous density (BIRADS B or C), with a statistically significant increase in the number of cancers detected in

heterogeneously dense breasts with DBT+s2DM compared to FFDM: 37 cancers detected with DBT+s2DM compared to 29 cancers detected with FFM only (21.6 percent; 95%CI: 11.4, 37.2; *p*=.039). Romero Martín et al. (2018) also reported that DBT+s2DM detected twice as many cancers in very dense breasts compared to the number of cancers detected with FFDM. Raw numbers however were very small (i.e., there were only two cancers in total in extremely dense breasts: one was detected with DBT+s2DM only and the other was detected with both DBT+s2DM and FFDM).

Results from the Verona trial were stratified by BIRADS A+B (not dense) and BIRADS C+D (dense). Statistically significant increases in CDR were reported for both groups of densities with DBT+s2DM compared to FFDM but the increase in dense breasts (BIRADS C+D) was much larger (12.9 cancers per 1000 screening exams when DBT+s2DM was used compared to 4.5 cancers per 1000 screening exams performed with FFDM, *p*=.001).

Retrospective analysis

A retrospective analysis (Rose & Shisler, 2018) reported statistically significant increases in CDR for women aged younger than 50 years and with more dense breasts, noting an adjusted incremental difference of 1.3 cancers detected per 1000 screening examinations (*p*=.039). Interestingly, while the CDR increased for women aged under 50 with more dense breasts when DBT was used, non-significant increases in CDR were detected for all women in the study. This may suggest that, for women aged under 50 years, DBT confers a benefit only when that woman also has dense breasts.

Conant et al. (2019) reported that CDR increased for all categories of breast density when FFDM+DBT was used. They reported breasts as non-dense (BIRADS 1 and 2) or dense (BIRADS 3 and 4). In this study, women with more dense breasts had a larger increase in detection when FFDM+DBT was used compared to women with less dense breasts (2.27 per 1000 screening examinations compared to 1.70 per 1000 screening examinations). Overall cancer detection increased for all women when FFDM was used. Information about the implications of breast density on recall rates from this study are reported in *section 3.4.1*.

Given the sample sizes, different imaging protocols and the different density strata used, more studies are needed to further explore the relationship between DBT (either with s2DM or FFDM), CDR and breast density to determine whether it is more beneficial for women with more dense breasts (and therefore may be appropriately considered as a more accurate test for some women). Reliability in stratification analysis by breast density is also likely to improve as systems to more consistently categorise density are used in research.

3.1.4. CDR stratified by age

Cancer incidence increases with age, and most breast-screening programs observe an increase in cancer detection with increasing age. Given that younger women are more likely to have more dense breasts compared to older women, and that use of DBT appears to increase CDR more in women with more dense breasts, we were interested to explore the literature to see whether there are also differences in CDR stratified by age. No systematic reviews of studies published since December 2017 or RCT data considered CDR stratified by age, but several trials and evaluation pilots did, reporting inconsistent findings. Inconsistencies may reflect the different ways in which CDR was stratified by age band. Detailed methodologies for each of these studies is provided in *section 3.1.1.*

Prospective studies embedded in population-based screening programs

CDR was stratified by age in the Maroondah and Verona trials, with both studies reporting increased cancer detection in older women compared to younger women for both screening tests (DBT+s2DM and FFDM alone). Significance was not achieved for the results for younger women. The studies also stratified by different age bands. In the Maroondah trial (Houssami et al., 2019), results were stratified by women aged over or under aged 60 years. For women aged over 60 years, DBT+s2DM detected more cancers (15 cancers detected per 1000 screening exams compared to 8.5 with FFDM; 95%CI: 1.1, 14). Non-significant results were presented for women aged under 60 years. The Verona trial stratified women into several age bands. For women aged 50-54 years, a non-significant increase in CDR was reported with the use of DBT+s2DM (6.8 per 1000 screening exams compared to 6.4 cancers detected with FFDM exams; p=.80). For women aged 55 years and over, CDR ranged between 7.8 and 13.0 per 1000 screening exams for DBT+s2DM compared to 3.4 to 6.1 cancers (FFDM), depending on the age band used. All comparisons by the older age band achieved significance (Caumo et al., 2018).

Zackrisson et al. (2018) reported on age stratified data from the Malmö trial (DBT_{MLO} compared to FFDM alone). For all age groups (40-49y; 50-59y; 60-69y), CDR increased with the biggest increases were seen in women aged 50-59 years:

- 40-49 years: DBT_{MLO} cancer detection was 19 percent compared to 16 percent
- 50-59 years: DBT_{MLO} cancer detection was 26 percent compared to 16 percent
- 60-69 years: DBT_{MLO} cancer detection was 43 percent compared to 36 percent
- 70-74 years: DBT_{MLO} cancer detection was 10 percent compared to 30 percent.

No statistical analysis or commentary on these results were provided by Zackrisson et al.

Retrospective analysis

Bahl et al. (2019) prepared a large retrospective analysis of the performance of FFDM+DBT to FFDM alone in women aged over 65 years. Participants in the FFDM+DBT group (n=20,646) had a mean age of 72.1 years; women in the FFDM (n=15,019) group had a mean age of 72.7 years. Bahl et al. reported no statistical difference in CDR between the two groups:

- FFDM+DBT: 8.2 cancers detected per 1000 screening examinations, compared to
- FFDM alone: 6.9 cancers detected per 1000 screening examinations (p=.23).

Given the limited available analysis exploring the relationship between older or younger age, use of DBT and CDR, more studies are needed to further to determine whether DBT is more useful at older ages or (as noted in Rose & Shisler's 2018 study) more useful in younger women with dense breasts.

3.1.5. Interval cancer rate

The interval cancer rate refers to the number of breast cancers that are diagnosed between screening examinations (i.e., 24 months in the BSA program). A true interval cancer cannot be seen on the previous mammogram (and may be a more aggressive cancer associated with a poorer health outcome); a missed cancer can be detected on imaging but was not. The interval cancer rate is an important long-term screening outcome and can be used as a surrogate indicator for screening benefit and is a marker of the sensitivity of a breast screening program. Improved

cancer detection due to better conspicuity can lead to a decline in interval cancers, especially missed cancers. Higher interval cancer rates may also be due to a longer screening interval, reader performance and interpretation failure, reading strategy, or may reflect an aggressive new cancer that was not previously visible on mammogram. If interval cancer rates decline, morbidity and mortality benefit may be inferred as aggressive cancers may be detected before they advance.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

At 31 December 2017, there was limited information about the impact of DBT (either alone or when used with FFDM) on interval cancer rate within population-based screening settings, mostly due to the study timeframes and powering (i.e., the studies looked at shorter-term outcomes like CDR/recall rates). None of the studies discussed reported statistically significant changes in interval cancer rates. For example, the STORM trial reported data on interval cancers (Houssami et al., 2014), which showed limited effect of DBT on reducing the interval cancer rate. 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 DBT. Non-significant results were also presented in a retrospective analysis from the PROSPR consortium. The previous GRADE assessment reflects the very limited evidence.

GRADE assessment: interval cancer: FFDM+DBT compared to FFDM alone

Participants Studies	Quality of evidence	Overall results
55,457 participants 3 studies	⊕ Very low	Few studies report interval cancer rates. No systematic review or pooled analysis was available.

GRADE assessment: interval cancer: DBT+s2DM compared to FFDM+DBT or FFDM alone

Participants Studies	Quality of evidence	Overall results
NA	Not reported	No data reported.

GRADE assessment: interval cancer: DBT_{MLO} compared to FFDM

Participants Studies	Quality of evidence	Overall results
NA	0	Nil

Updated findings

Literature published since 31 December 2017 reported results for the Malmö, OTS and STORM trials. Whether early recall is practiced if an inconclusive screening result is recorded and the way that interval cancer rate is defined and reported in these studies is important, as there is definition variability within studies (described below), which makes comparison between studies and to the BSA program difficult. In addition, more recent larger studies are powered to compare interval cancer rates; however, these results have not yet been reported as further follow-up time is required. No systematic reviews of studies published since December 2017 have considered

interval cancer rate. Included studies are listed below. Detailed methodologies for each of these studies is provided in *section 3.1.1.* Included studies are listed below.

Literature review

One narrative literature review: Chong et al., 2019

Prospective studies embedded in population screening programs

Four studies: Bernardi et al., 2019; Houssami et al., 2018; Skaane et al., 2018; Zackrisson et al., 2018

Observational studies

One retrospective analysis: Hovda et al., 2019

Summary of key findings published between 1 January 2018 and 31 December 2019

Interval cancer rate is an important long-term screening outcome, especially as interval cancers are often more advanced and can be associated with higher morbidity and mortality compared to screen-detected cancers. Several pilot evaluation or paired studies have reported on interval cancer rates; however, reported results have generally come from studies that have not been powered to calculate interval cancer rates or which have varying screen intervals (as per the program's policy settings). Other studies like the Reggio Emilia RCT are powered to detect interval cancers but results have not yet been reported.

Current data is indicative only. Adequate data evaluating interval cancer rate by imaging modality is not yet available. Reported results vary and no study has reported a statistically significant reduction in interval cancers with the use of DBT. Few studies compared interval cancer characteristics by type, grade, node status and size, which would be helpful to assess whether DBT detects more aggressive cancers earlier. Only the OTS trial has reported this and found no significant differences the types of cancers detected between pre- and post- screening with DBT. This suggests that DBT may not influence the detection of more aggressive cancers earlier; however, at this stage, we do not know the impact of DBT on decreasing interval cancer rates or whether interval cancers differ in terms of pathology/histology from cancers detected within in a screening round or with interval cancers from previous screening rounds. Determining DBT's influence on interval cancers remains one of the key areas to explore in future research. The screening interval (i.e., annual or biennial) and the way that interval cancer rate is defined and reported in studies is important as this could also account for variability within study results. Pooled analysis work, which will include data from the main prospective paired studies set in population-based screening programs, is underway and due to report in 2020.

Literature review

Chong et al. (2019) reported data presented from the OTS and STORM trials (see below for further discussion of these trial results) and retrospective American studies which reported 1.1 per 1000 screens interval cancer rate for both imaging protocols (Bahl et al., 2018). Study results vary but no results achieve statistical significance: STORM data indicated a slight non-significant reduction in interval cancer rates and the OTS trial and the Bahl et al. studies showed no statistically

significant difference in interval cancer rate when DBT is used. We note that there are a number of important considerations here, notably that Bahl et al.'s study used annual screening (and is likely to have a lower interval cancer rate compared to biennial screening) whereas the other studies are biennial. Chong et al. concluded that more information is needed before understanding the influence of DBT on reducing interval cancer rates is reached. Research to address this gap is underway: Houssami et al. (2017) are undertaking an individual participant data-analysis of prospective trials set in biennial population-based screening programs to compare DBT (either alone or integrated with FFDM or s2DM) to FFDM. Included studies must have published data on interval cancer rates and at least two-years follow-up. Analysis on interval cancer results is due in 2020 and will progress our understanding of the impact of DBT on interval cancer rates.

DBT+s2DM compared to FFDM+DBT and/or FFDM alone

Prospective studies embedded in screening programs

Study Design	Results: interval cancer rate (95%CI)
Bernardi et al. (2019) Trento pilot evaluation (STORM site)	At two years follow-up: DBT+s2DM: 1.1 FFDM alone: 1.36 RR=.81 (0.55, 1.19)

Data from the Trento pilot evaluation reported on interval cancer rate at two years follow-up (Bernardi et al., 2019). The authors noted that 51 interval cancers were detected in each cohort (i.e., women screened with DBT+s2DM between October 2014 and October 2016 compared to women screened with FFDM only between January 2013 and October 2014). The interval cancer rate was slightly lower for women screened with DBT+s2DM (see box above) but the relative risk was not statistically significant. The authors note that it is likely that this study was not powered to detect significant changes in interval cancer rates. No information was provided about the characteristics of the interval cancers detected in this pilot. Six interval cancers were identified in the STORM-2 trial (Bernardi et al., 2018) but no further comment or analysis on this finding was made.

The Reggio Emilia trial (Pattacini et al., 2018, methodology described in *section 3.1.1*) is powered to compare interval cancer rates, but it has not yet reported results.

Retrospective analysis

Hovda et al. (2019) completed a retrospective analysis of data from the Norway BreastScreen Program. The authors compared interval cancer rates using paired data from women with two or more consecutive screening examinations. A slightly different interval cancer definition was used by Hovda et al.: breast cancer diagnosed after a negative mammogram result or within six months of a false positive result and both within two years of a screening examination. Like Skaane et al. (2018), Hovda et al. reported no statistically significant differences in interval cancer regardless of the screening examination undertaken:

- Interval cancer rate for Group 1 women (two FFDM screens): 2.2 per 1000 women screened
- Interval cancer rate for Group 2 women (FFDM then FFDM+DBT or DBT+s2DM): 1.9 per 1000 women screened.

They reported on specific histological characteristics of interval cancers, finding statistically significant decreases in the number of grade 1 intervals cancers when women were screened with DBT (either FFDM+DBT or DBT+s2DM) after a previous FFDM examination: 0.6 grade 1 interval

cancers were detected per 1000 screening examinations compared to 0.67 with FFDM (p<.01). more grade 3 interval cancers were reported with DBT, but this result was not significant. No other statistically significant results were reported.

FFDM+DBT compared to FFDM alone

Prospective studies embedded in screening programs

Two trials set in population-based screening programs (OTS and STORM trials) have reported on interval cancer rates, but neither reported statistically significant increases in interval cancer following use of DBT (see box, right). Skaane et al. (2018) reported two-year follow-up data on interval cancer rate from the OTS trial, noting few differences following use post screening with FFDM+DBT (methodology described in *section 3.1.1*). In this study, the authors compared 51 interval cancer rates from the OTS trial with aggregate screening data from two previous FFDM screening rounds. The interval cancer analysis was not paired. The OTS interval cancer rate was 2.1 interval cancers per 1000 screening examinations compared to 2.0 interval cancers per 1000 screening examinations in the previous two screening rounds. The incremental difference was not statistically significant. Interestingly, while the interval cancer rate was not different, the authors noted that proportionally more interval cancers presented after one year when FFDM was used (73 percent compared to 61 percent) but did not provide further comment of the importance of this finding (or otherwise).

Study Design	Results: interval cancer rate Rate per 1000 women screened (95%CI)
Houssami et al. (2018) STORM	At two years follow-up: FFDM+DBT: 1.23 (0.56, 2.34) FFDM alone: 1.60 (1.14, 2.17)
Skaane et al. (2018) OTS	FFDM+DBT: 2.1 FFDM alone: 2.0 Incremental difference: 0.1 (-0.5, 0.8, p=.734)

STORM was not powered to detect interval cancer rates; however, Houssami et al. (2018) completed reporting of the STORM trial results with a paper on interval cancer rates. In this trial, 7292 asymptomatic, average-risk Italian women aged 48 years or older (median age 58 years) were screened with FFDM+DBT, with reading of FFDM images only followed by FFDM+DBT reading (i.e., a paired trial). Interval cancer was defined as a cancer identified after a negative mammogram before the next routine screening exam and included DCIS. This final STORM paper reported on interval cancers with a minimum follow-up of 24 months from the date of the screening exam (i.e., this study builds on Houssami et al.'s 2014 results, which had 13 months follow-up). A total of nine interval cancers were detected (three of which were detected in the first 12 months of follow-up, which is a slightly different rate compared to what was reported in the 2014 paper). Non-significant rate results were 1.23 interval cancers per 1000 screening examinations (95%CI: 0.56, 2.34) when DBT was used or 1.60 per 1000 screening examinations (95%CI: 1.14, 2.17) when FFDM alone was used.

Skaane et al. (2018) and Houssami et al. (2018) reported on interval cancer characteristics. Skaane et al. compared characteristics between post FFDM+DBT interval cancers compared to interval cancers from the two previous FFDM alone rounds. Houssami et al. reported only on the interval cancers detected in STORM.

In the OTS trial, Skaane et al. reported that the interval cancers were similar between screening rounds; that is, there was no statistical difference in characteristics by interval cancers detected post-screening with FFDM+DBT compared to FFDM alone. Specific findings include that DBT-

detected interval cancers were lower grade, less likely to be node positive, and smaller than cancers detected with FFDM. Results are summarised in the bullets below:

- Cancer type: most interval cancers were invasive with slightly more DCIS detected with FFDM alone (FFDM: 95.8 percent compared to 96.1 percent post-FFDM+DBT), which is similar to the results for cancers detected at screening
- Molecular subtype: interval cancers only detected with DBT were mostly luminal A or luminal B HER-2 negative cancers with low Ki67 (i.e., cancers which have a good prognosis even if detected at a subsequent screening round), which is similar to the results for cancers detected at screening
- Grade: most interval cancers were grade 2 or 3 (FFDM: 41.1 percent and 44.9 percent compared to 41.3 percent and 47.8 percent post-FFDM+DBT, grades 2 and 3 respectively); interval cancers only detected with DBT were lower grade than those detected in the previous FFDM alone rounds (i.e., the difference in Grade 1 invasive cancers was 35 percent more with DBT, $p<.001$), which is similar to the results for cancers detected at screening
- Nodal status: most interval cancers were node negative and the results were similar between imaging protocols (FFDM: 60.9 percent compared to 60.4 percent post-FFDM+DBT); interval cancers only detected with DBT were much more likely to be node negative (96.1 percent of cancers detected only with DBT were node negative compared to 60.9 percent detected in the two previous FFDM rounds, $p<.001$), which is similar to the results for cancers detected at screening, and
- Size: interval cancers detected only with DBT were more likely to be smaller than ≤10mm compared to those detected in the previous two screening rounds using FFDM (51 percent compared to 17.2 percent, $p<.001$), which is similar to the results for cancers detected at screening.

All this data suggests that DBT may not result in the earlier detection of more aggressive cancers compared to FFDM alone.

Houssami et al. (2018) reported on cancer characteristics for the nine interval cancers identified in STORM. Seven cancers were invasive ductal cancers, one was an invasive lobular cancer and one was DCIS (which is similar proportionally to the types of interval cancer detected in the OTS trial). In STORM, interval cancers were generally advanced (i.e., seven tumours were larger than 20mm, five had a Ki-67 above 24, and five were grade 3 although only one cancer had nodal metastases), which (as noted by the authors) is in keeping with the types of interval cancers detected in other biennial screening programs. Houssami et al. makes no comment on whether the interval cancers differed from the screen-detected cancers in the STORM trial.

DBT_{MLO} compared to FFDM

Prospective studies embedded in screening programs

Interval cancer data from the Malmö trial (methodology described in *section 3.1.1*) was reported by Zackrisson et al. (2018). Follow-up times differed by age: women aged 40-54 years were re-screened at 18-month intervals, women aged 55-74 years were screened at two-year intervals. The authors reported 22 interval cancers and a study population interval cancer rate of 1.48 cancers per 1000 women screened. No data comparing the rates of interval cancer between

DBT_{MLO} and FFDM, or between interval cancers detected before/after the use of DBT was provided in this paper, but further analyses are planned. Zackrisson et al. did note that interval cancers compared to screen-detected cancers tended to be invasive and were more advanced. About 90 percent of the interval cancers were invasive (mostly IDC), medium-sized (mean tumour size 17mm), five had nodal metastases and five were grade 3 tumours. An interesting finding was that eight cancers were only detected by FFDM but were noted during the DBT reading but none of these women were recalled as the findings were not considered sufficiently suspicious on DBT (presenting as microcalcifications). The authors however do not comment on whether these could have become interval cancers/missed cancers if only the DBT reading protocol was followed.

3.1.6. Relative sensitivity

The sensitivity of a screening test refers to the proportion of breast cancers correctly identified by a test or the true positive rate. Increased sensitivity with DBT (compared to FFDM) indicates that use of DBT correctly identifies more breast cancers.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

Because follow-up and overall study timeframes in the primary studies published to 31 December 2017 tended to be shorter, few estimates of absolute sensitivity for FFDM+DBT are available, but there was some data reporting on the *relative* sensitivity of FFDM+DBT as a screening test compared to FFDM alone, DBT+s2DM as a screening test compared to FFDM or DBT_{MLO} compared to other imaging combinations. Mixed results were identified. Some studies (including data from trials embedded in population-based screening programs) reported large increases in relative sensitivity with increases in all readers' performance when DBT was included in the imaging pathway. Some smaller studies using enriched samples reported no statistically significant change. No GRADE assessment was completed for this outcome due to lack of data.

Results from systematic review studies included in the literature review on the role of DBT in the assessment of screen-detected abnormalities reported statistically significant increases in sensitivity with the use of DBT although we note that some of the results were drawn from studies of diagnostic populations. That said, for DBT's role in assessment, pooled analysis of seven studies using histological results as the reference standard showed a pooled sensitivity of 90 percent (95%CI: 87, 92) for DBT compared to 89 percent (95%CI: 86, 91). Phi et al. (2018) reported that, in diagnostic populations, DBT had higher sensitivity compared to DM for women with dense breasts: 84 to 89 percent for DBT compared to 69 to 86 percent for DM images.

Updated findings

Literature published since 31 December 2017 provides a little more (but not much) information on sensitivity compared to what was available in the 2017 literature review on DBT's sensitivity as a screening test. Most of the research discusses sensitivity in relation to the diagnostic performance of DBT within the assessment clinic (i.e., as part of investigation of a screen-detected abnormality) rather than in relation to the screening examination itself. As such, this section of the report is organised differently with only summary results presented. Studies that discussed sensitivity in relation to screening or assessment of screen-detected abnormalities are listed

below. No systematic reviews of studies published since December 2017 have considered program sensitivity. Detailed methodologies for each of these studies is provided in *section 3.1.1.*

Prospective studies embedded in population screening programs

Four studies (five papers): Bernardi et al., 2019; Skaane et al., 2019; Houssami et al., 2018; Skaane et al., 2018; Zackrisson et al., 2018

Observational studies

Three studies with a retrospective design: Bahl et al., 2019; Conant et al., 2019; Hovda et al., 2019

Summary of key findings published between 1 January 2018 and 31 December 2019

All studies reported increases in sensitivity when DBT was used but non-significant results or wide confidence intervals were also noted. Increases were observed with all reading protocols (i.e., FFDM+DBT, DBT+s2DM). When increases in sensitivity were observed, the increase ranged from less than one percent more to over 16 percent more. Larger increases in sensitivity were observed in the screening programs most similar to the BSA program (double read, biennial screening).

DBT+s2DM compared to FFDM (alone or with DBT)

- Trento pilot evaluation: 88.74 percent with DBT+s2DM compared to 80.08 percent (Bernardi et al., 2019).
- OTS: 69.0 percent with DBT+s2DM compared to 70.5 percent wither FFDM+DBT (Skaane et al., 2019).
- Hovda et al.'s retrospective analysis of Norway BreastScreen Program data: sensitivity increased when DBT+s2DM or FFDM+DBT was used subsequent to a FFDM round (83.7 percent) compared to when FFDM was used for both screening rounds (67.6 percent).

FFDM+DBT compared to FFDM alone

- OTS trial: double reading of FFDM+DBT increased sensitivity compared to FFDM (80.8 percent with FFDM+DBT compared to 76.2 percent in the previous two screening rounds; +4.6 percent; 95%CI: -1.4, 10.5, p=.151) (Skaane et al., 2018).
- OTS trial: reader adjusted sensitivity also favoured single reading of FFDM+DBT over a single reading of FFDM alone (70.5 percent compared to 54.1 percent) and double reading of the FFDM+DBT and DBT+s2DM had a higher sensitivity (83.3 percent) compared to double reading of FFDM alone and FFDM+CAD (65.5 percent) (Skaane et al., 2019).
- STORM: use of FFDM+DBT increased sensitivity compared to FFDM: 85.5 percent (95%CI: 75.0, 92.8) with FFDM+DBT compared to 77.3 percent (95%CI: 70.4, 83.2) with FFDM alone (Houssami et al., 2018).

- A retrospective analysis (Bahl et al., 2019): FFDM+DBT increased sensitivity to 90.9 percent (95%CI: 87.5, 93.4) compared to FFDM alone (88.3 percent; 95%CI: 83.6, 91.8, *p*=.39).
- A retrospective analysis (Conant et al., 2019): sensitivity for FFDM was 91.5 percent compared to 90.6 percent for DBT for all women, which is counter-intuitive given the higher CDR observed when DBT was used in this study: the authors noted that other imaging was also used (including ultrasound and MRI which may have increased overall program sensitivity for the FFDM arm and as such is unlikely to reflect a true comparison between FFDM+DBT and FFDM alone).

DBT_{MLO} compared to FFDM alone

- Malmö trial: use of DBT_{MLO} increased sensitivity compared to FFDM (81.1 percent; 95%CI: 74.2, 86.9 with DBT_{MLO} compared to 60.4 percent; 95%CI: 52.3, 68) (Zackrisson et al., 2018).

3.2. Cancer type and histopathological and prognostic/predictive tumour characteristics

There is consistent (but not unequivocal) evidence that use of DBT increases overall CDR (see *section 3.1*). Breast screening is intended to reduce mortality from breast cancer via early detection. While it is important to know whether an imaging technique accurately depicts cancers, it is also important to understand the characteristics of that cancer given the heterogeneous nature of breast disease. Some cancers have a favourable prognosis even when detected at a fairly large size compared to more aggressive cancers with a poorer prognosis. If use of DBT results in detection of more clinically significant invasive cancers at an earlier stage, its use may confer an additional morbidity and/or mortality benefit over FFDM alone (i.e., it detects 'killing' cancers earlier and enables more aggressive treatment or greater treatment options). If more ductal carcinoma in situ (DCIS) or low-grade, slow-growing invasive cancers are detected, the morbidity or mortality benefit associated with detecting and treating these small cancers may be lower. Alternatively, slow-growing cancers may be detected earlier but the favourability of outcome prognosis may not change compared to detecting the cancer at a later date. Key dimensions that are used to assess this relationship are:

- The type of cancer (i.e., invasive cancers or DCIS) and the molecular sub-type
- Tumour staging (the histological grade, tumour size and lymph node involvement), and
- Radiological presentation.

We want to know, based on current evidence, whether DBT detects clinically significant cancers, or whether the additional cancers detected with DBT are not significant and may represent overdiagnosis. Papers covered in this literature review reported on the type of cancer and varying dimensions of tumour staging but few reported on radiological presentation. In this section, we explore the current evidence base on whether cancers detected with DBT are likely to be clinically important or whether they are abnormalities that may not likely to become life-threatening or symptomatic within a woman's lifetime (i.e., they may contribute to overdiagnosis).

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

There was some evidence that detection of invasive cancers increased when using DBT compared to FFDM alone. Increases were reported in both screening and assessment settings and for different combinations of screening strategies (i.e., FFDM+DBT, DBT+s2DM, and DBT_{MLO} compared to FFDM alone).

Early evidence from the primary screening studies indicated that DBT may detect smaller cancers which are more likely to be invasive. A fixed effect meta-analysis of data from the STORM and OTS trials (Hodgson et al., 2016) found a statistically significant increase in invasive CDR detected by FFDM+DBT compared to FFDM alone (i.e., an additional 2.33 invasive cancers per 1000 screening examinations). Interim data from STORM and Malmö trials reported that DBT appears to detect microcalcifications as well as FFDM. 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. Retrospective studies also reported that invasive CDR increased with the use of FFDM+DBT but not all results achieved statistical significance. Bernardi & 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). Findings regarding detected DCIS were mixed, but no statistically significant results suggested that DBT increased detection of DCIS and there was insufficient evidence to describe the relationship between DBT and overdiagnosis, as few studies reported on specific cancer types and/or characteristics (including no studies that reported on molecular subtype). The literature review concluded that more research was needed to understand the differences between cancers detected with/without DBT.

In the assessment setting (and based on studies which included screening populations, mixed populations, and diagnostic populations alone) DBT, alone or in combination with FFDM, detected significantly more invasive cancers compared to FFDM. Key results included a RR=1.327 for the increase of invasive cancer with FFDM+DBT compared to FFDM and significant increases in detection of invasive ductal carcinoma (RR=1.437) and special type carcinomas. DBT detected more T1 cancer (RR=1.388) and T1NO cancers. T1N0 cancers were smaller and are likely to not have spread to auxiliary lymph nodes, making them clinically important as these cancers detected at screening are more likely to have a better prognosis. DBT's benefit in detecting cancers T2 or larger was less certain and no consistent differences in detection were reported for Grade 2 or 3 cancers. There were mixed results for the detection of node negative disease and by molecular subtype (although there was limited evidence on these characteristics).

GRADE assessment: invasive CDR: FFDM+DBT compared to FFDM alone

Participants Studies	Quality of evidence	Overall results
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.

GRADE assessment: invasive CDR: DBT+s2DM compared to FFDM+DBT or FFDM alone

Participants Studies	Quality of evidence	Overall results
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.

GRADE assessment: invasive CDR: DBT_{MLO} compared to FFDM

Participants Studies	Quality of evidence	Overall results
7,681 2 studies	⊕ Very low	No systematic review or pooled analysis was available.

Updated findings

The To-Be-1 RCT and most of the prospective studies embedded in population-based screening programs reported on at least some cancer characteristics (notably cancer type, grade, size and lymph node involvement). Only a small number of studies reported on molecular sub-type.

Most studies provide descriptive results by imaging protocol. In most of the study designs[13], cancers were analysed by whether they were detected in the DBT arm or the study or the FFDM arm (i.e., cancers detected in the DBT arm are not 'additional' cancers to those detected in the FFDM arm). It makes it more challenging to determine cancers that are only detected with DBT are different from those detected with DBT+s2DM or FFDM. Few studies included statistical analysis. This is not a criticism of these studies but a reflection that the number of cancers detected in most studies was small and studies were not sufficiently powered for robust analysis by cancer type and characteristics. No systematic reviews of studies published since December 2017 have considered cancer type or tumour characteristics. Included studies are listed below. *Tables 10 – 14* include a summary of all study results by key characteristics.

Randomized controlled trials

One RCT (one paper): Hofvind et al., 2019

One randomized study embedded in a population screening program: Pattacini et al., 2018

[13] This includes the To-Be-1 RCT, Maroond1ah

Prospective studies embedded in population screening programs

Seven studies (eight papers): Bernardi et al., 2019; Houssami et al., 2019; Johnson et al., 2019; Caumo et al., 2018; Hofvind et al., 2018; Romero Martín et al., 2018; Skaane et al., 2018; Zackrisson et al., 2018

Observational studies

Six studies with a retrospective design: Bahl et al., 2019; Conant et al., 2019; Dang et al., 2019; Fuiji et al., 2019; Hovda et al., 2019; Rose & Shisler 2018

Summary of key findings published between 1 January 2018 and 31 December 2019

Overall, use of DBT appears to detect more invasive disease than FFDM (including when DBT is used in combination with FFDM or s2DM or if DBT_{MLO} is used alone). Evidence describing whether DBT detects more DCIS is inconsistent: most study results achieving statistical significance suggested that more DCIS is detected with DBT, but more studies suggest there is no increase or less detection with DCIS. More research is needed to fully unpick these results. It is also clear that DBT+s2DM does not result in less detection of either invasive cancers or DCIS (i.e., s2DM can probably be used instead of FFDM, reducing radiation dose).

Information relating to molecular subtype, grade, size and lymph node status is also important in assessing whether cancers detected are clinically significant. While data was grouped in different bands making it difficult to fully assess this relationship, there is sufficient evidence to suggest that cancers detected with DBT are likely to be slightly smaller, lower grade and node-negative (i.e., potentially earlier stage cancer) than those detected with FFDM. Most studies reported no or only small differences in the proportion of node negative and node positive disease, regardless of the imaging used (i.e., DBT+s2DM, FFDM+DBT or DBT_{MLO}). If there was a difference, it usually favoured the DBT imaging arm (i.e., DBT detected more node-negative disease). Mixed results were also presented for grade, with some studies suggesting that DBT detected more low-grade cancers compared to those detected with FFDM.

Generally, it appears that the cancers detected by DBT are similar in terms of histological type to those detected with FFDM (although more cancers presenting as spiculated masses and architectural distortion are detected with DBT). Across all studies (regardless of the way DBT was used), increases in IDC were usually observed with the use of DBT. Mixed results were presented for ILC, with some studies demonstrating a statistically significant increase in ILC detection with FFDM, but others noting an increase in ILC detection when DBT was used (including results from the Malmö trial). Mixed results were also presented about molecular subtype. Some studies (including Hofvind et al.'s 2018 and Hovda et al.'s 2019 analysis of data from the Norway BreastScreen Program and baseline data from Pattacini et al.'s 2018 RCT) showed an increase in cancers with a molecular profile more favourable to good prognosis (i.e., more Luminal A cancers, more Luminal B Her-2 negative cancers and more cancers with Ki67%$\leq$20). Others, including data from the To-Be-1 RCT and the Malmö trial showed no significant differences in molecular subtype or receptor status.

Some of these differences may be due to insufficient powering, descriptive rather than comparative analysis, or the way that cancers are grouped according to the imaging in which they were detected. There is not yet sufficient evidence to determine if additional cancers detected with DBT alone are more aggressive 'killing' cancers, if they are comparable to the

types of cancers detected with FFDM (but more are detected) or if additional cancers detected may add to the overdiagnosis burden; however, there is accumulating evidence that use of DBT does not contribute to an increase in detection of DCIS (although it could still contribute to overdiagnosis through the detection of less aggressive cancers like Luminal A cancer).

3.2.1. Histological type

Invasive cancer detection rate refers to the total number of invasive cancers that can be identified using a specific imaging technique(s). In the literature, most studies only distinguished between invasive cancers and DCIS, although some data from the prospective studies embedded in population-based screening programs also talked about specific histological types like invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC) and other special-type cancers like papillary, tubal and mucinous carcinomas (see *Tables 9a* to *9c* for overall data on invasive CDR by imaging protocol and study, and *Table 10* for information about specific histological type where this was reported). This additional detail is important as some types of cancer have a worse prognosis than others, and earlier detection of these 'killing' cancers may result more treatment options for the women and a commensurate reduction in mortality for the program.

Across all studies (regardless of the way DBT was used), increases in IDC were usually observed with the use of DBT. Mixed results were presented for ILC, with some studies demonstrating a statistically significant increase in ILC detection with FFDM, but others noting an increase in ILC detection when was DBT used (including results from the Malmö trial).

DBT+s2DM compared to FFDM alone

Most population-based studies embedded in screening programs reported that use of DBT+s2DM detected more invasive cancers and the same amount or less DCIS compared to cancers detected in the FFDM arm (Bernardi et al., 2019; Caumo et al., 2018; Romero Martín et al., 2018). Some studies reported an increase in DCIS detected as well when DBT+s2DM was used (Houssami et al., 2019; Hofvind et al., 2018). Increased detection of invasive cancers is consistent with *Allen + Clarke*'s previous two literature reviews; however, some studies still reported inconsistent results (including results from the To-Be-1 RCT which showed no statistically significant increases). It is also important to understand whether the increase in invasive cancers represented an increase in those with a more aggressive or worse prognosis. Information relating to molecular subtype, grade, size and lymph node status is also important.

RCT

The To-Be-1 RCT (which compared DBT+s2DM to FFDM alone) reported a non-significant increase in invasive CDR (DBT+s2DM detected 5.6 invasive cancers per 1000 screening examinations; FFDM detected 4.9 cancers per 1000 screening examination) and DCIS CDR (DBT+s2DM detected 5.6 DCIS per 1000 screening examinations; FFDM detected 4.9 DCIS per 1000 screening examination). No differences were observed by cancer type or molecular subtype, grade, mean tumour size or lymph node status between cancers detected with the DBT or FFDM arms (Hofvind et al., 2019). Hofvind et al.'s paper was also the only prospective study to separately report on DCIS size and grade, but they did not report any significant differences in size (although 50 percent of the DCIS detected with DBT+s2DM was ≥20mm compared to 76.9 percent of the DCIS detected with FFDM, which suggested that the DBT arm may detect more smaller DCIS).

Prospective studies embedded in screening programs

Summaries of the main trials/evaluations set in population-based screening programs are summarised below:

- Cordoba (Romero Martín et al., 2018): use of DBT+s2DM resulted in a statistically significant increase in the detection of invasive cancers, most of which were low-grade cancers ≤10mm with statistically significant increases in IDC detection (but not other cancer types and no molecular subtyping provided) and no statistically significant increase in detection of DCIS.
- Maroondah (Houssami et al., 2019): using DBT+s2DM resulted in the detection of more low-grade, node negative invasive cancers (especially IDC and special-type cancers and HER-2 positive cancers) and low to intermediate grade DCIS compared to FFDM alone but no statistical analysis was completed due to the small number of cancers (n=49 detected in the DBT arm and 34 in the FFDM alone arm).
- Norway BreastScreen Program analysis:
  - Hofvind et al. (2018): more grade 1 invasive cancers ≤10mm (including statistically significant increases in IDC and some special cancers like tubular cancers) and an increase in all molecular subtypes with use of DBT (including subtypes that have a good prognosis or which may not become clinically significant in a woman's lifetime) and increased DCIS detection.
  - Hovda et al. (2019): their retrospective analysis reported non-significant but higher rates of invasive cancers detected with DBT+s2DM compared to FFDM. This included more small (≤20mm), grade 1 and grade 2 invasive cancers (<20mm), including cancers with a molecular subtype with good prognosis (Luminal A, Luminal B HER-2 negative for women undergoing DBT+s2DM after an earlier FFDM screen) (i.e., cancers that may progress clinically are being detected earlier with the use of DBT), and
- Trento (Bernardi et al., 2019): DBT+s2DM detected slightly more Grade 1, node negative, stage 1 and 2 invasive cancers (particularly those ≤10mm and stage I and II) compared to FFDM alone without an increase in DCIS detection.
- Verona (Caumo et al., 2018): DBT+s2DM detected more smaller invasive cancers (mean=15.1mm compared to 18.6mm), less pT2+ cancers, and less DCIS compared to FFDM alone but statistical significance was not achieved for many of these cancer characteristics (including the staging findings).

FFDM+DBT compared to FFDM alone

Results from the trials comparing FFDM+DBT to FFDM alone reported that use of DBT increased detection of some types of invasive cancers as well as detecting more DCIS compared to FFDM alone. Not all results indicating an increase in DCIS detection were significant (and several studies reported no differences in terms of DCIS detection with FFDM+DBT).

RCT

In the Reggio Emilia RCT (Pattacini et al., 2018), there was significant differences in the number of grade 2 invasive cancers (and in cancers with PR negative, ER negative and HER-2 positive molecular subtypes) detected with DBT+s2DM compared to FFDM as well as more DCIS was detected with DBT+s2DM (although the authors noted that the increase of +1 DCIS case per 1000 screening examinations was within the range of DCIS detection needed to identify progressing invasive disease). Ki67≤20 percent was higher when DBT+s2DM was used, suggesting potential for DBT to identify less aggressive tumours but the authors concluded that, taken together, the weight of evidence suggested earlier diagnosis of invasive cancers.

Prospective studies embedded in screening programs

Final data from the OTS trial (Skaane et al., 2018) reported that FFDM+DBT detected more invasive cancers (including IDC and Luminal A subtype) with more node negative disease, and less DCIS but no difference in tumour size grade than was detected in the two previous FFDM rounds, suggesting the earlier detection of invasive cancers with a better prognosis.

Retrospective analysis

Bahl et al. (2019) reported on their matched patient cohort analysis and observed that CDR for invasive cancers was higher when DBT was used as part of the screening examination: an invasive CDR of 2.8 invasive cancers per 1000 screening examinations was reported for DBT imaging compared to 1.3 invasive cancers with FFDM (RR=2.2; 95%CI: 1.2, 3.9, p=.01). No statistically significant difference in DCIS (defined in this study as non-minimal cancers) was reported. No information on cancer type was reported in this study (beyond 'invasive').

DBT_{MLO} compared to FFDM alone

Prospective studies embedded in screening programs

Two papers reported results about tumour characteristics and molecular subtypes for the Malmö trial (Johnson et al., 2019; Zackrisson et al., 2018). Overall, DBT_{MLO} detected more invasive cancers than FFDM (87 percent of cancers detected were invasive compared to 82 percent with FFDM). Slightly less DCIS was detected with DBT_{MLO} (13 percent compared to 18 percent). In this trial, the DBT_{MLO} arm detected more early-stage invasive, Luminal A cancers compared to those detected in the FFDM arm, with slightly more ILC but there was no difference in the grade, tumour size or nodal involvement, or in molecular subtype (although more triple negative cancers were detected only by DBT). While no statistically significant increase in ILC was observed, the authors reported that additional detection of these cancers could be clinically important as they are harder to see on FFDM and increased conspicuity on DBT_{MLO} could result in earlier detection. The DBT arm resulted in detection of slightly less DCIS. Johnson et al. concluded that more Luminal A cancers were detected with DBT (51 percent compared to 46 percent; no significance testing) and this could contribute to overdiagnosis (rather than overdiagnosis from additional DCIS detection).

Table 10: Histological type for a range of imaging protocols

Study	Cancer type (%)							
	IDC		ILC		Specials[14]		DCIS	
	DBT arm	**FFDM arm**	**DBT arm**	**FFDM arm**	**DBT arm**	**FFDM arm**	**DBT arm**	**FFDM arm**
DBT+s2DM compared to FFDM alone								
RCT								
Hofvind et al. (2019) To-Be-1 (main analysis)	77.5%	71.8% (*p*=.11)	7.5%	18.3%	Tubular: 2.5% Other: 12.5%	Tubular: 5.6% Other: 4.2%	1.0 DCIS per 1000 screens	1.1 DCIS per 1000 screens
Prospective studies embedded in population-based screening								
Houssami et al. (2019) Maroondah	75%	70%	13%	23%	13%	7%	9%	4%
Hofvind et al. (2018) Norway BSP	5.6 per 1000	4.9 (*p*=.014)	1.1 per 1000	0.9 per 1000 (*p*=.061)	0.9 per 1000	0.3 for tubular carcinoma per 1000 (*p*<.001)	1.7 per 1000	0.8 (*p*<.001)
Romero Martín et al. (2018), Cordoba	67 cancers	55 cancers (*p*=.001)	4 cancers	2 cancers (*p*=.102)	No data	No data	21 cancers	19 cancers (*p*=.774)
Retrospective analysis								
Hovda et al. (2019) Norway BSP	6.43 per 1000	2.48 per 1000, (*p*=.001)	0.94 per 1000	1.14 per 1000	Tubular: 0.74 per 1000, (*p*=.01) Other: 0.24 per 1000	Tubular: 0 Other: 0.10 per 1000	1.6 per 1000	0.9 per 1000 (*p*=.001)
FFDM+DBT compared to FFDM alone								
Retrospective analysis								
Dang et al. (2019)	75%	72%	No data	No data	No data	No data	25%	28%
Fuiji et al. (2018)[15] Vermont	265 cancers	335 cancers	42 cancers	42 cancers	26 mixed cancers 12 other cancers	24 other cancers 30 mixed cancers	91 cancers	117 cancers
DBT_{MLO} compared to FFDM alone								
Prospective studies embedded in population-based screening								
Johnson et al. (2019) Malmö[16]	57%	67%	30%	17% (*p*=.13)	14%	15%	8%	18%
Zackrisson et al. (2018) Malmö[18]	52%	55%	26%	14%	12%	12% (tubular) 1 (other)	10%	18%

[14] This includes tubular, cribriform and mucinous types (Maroondah); tubular and other not described (Malmö)
[15] None of the OR calculated for the differences between cancer type by screening protocol were significant.
[16] Figures are the number of additional cancers detected with DBT_{MLO} compared to the number of cancers detected in the DM arm (i.e., the cancers only detected by DBT compared to the cancers detected by FFDM).

While most of these studies suggest that DBT increases detection of invasive cancers compared to in-situ cancers, many studies reported descriptively rather than providing a comparative statistical analysis of the additional cancers only detected with DBT due to the overall small number of cancers detected in studies. Also, it is unclear whether the cancers detected represent additional and earlier detection of clinically significant cancers (which would favour use of DBT as a screening test), or whether the cancers detected in the DBT arms of the study represent an increased lead time (i.e., these cancers would have been detected with FFDM eventually and they are the type of cancer than has a good prognosis). While evidence continues to emerge that DBT may increase the detection of invasive cancers (and not DCIS), the contribution of DBT to potential overdiagnosis or increased lead time (rather than increased detection of 'killing' cancers) requires additional research.

3.2.2. Molecular subtype

Approximately 70 percent of breast cancers are sensitive to specific hormones, meaning that presence of receptor-positive cells in a breast cancer may result in faster proliferation of cancerous cells. Molecular subtypes are based on the genes expressed by a cancer and are important biomarkers used to determine prognosis following diagnosis. There are four key biomarkers based on different configurations of receptors[17]:

- **Luminal A** breast cancer (ER positive, PR positive, HER-2 negative, low Ki67[18]) is usually associated with low-grade, slow-growing cancers with an excellent prognosis
- **Luminal B** breast cancer (ER positive and/or PR positive, HER-2 negative or positive, high Ki67) is usually associate with slightly faster growing cancers (compared to Luminal A)
- **Triple negative** breast cancer (ER negative, PR negative, HER-2 negative) is usually an aggressive cancer, is often associated with BRCA1 mutations and in younger women: it has poor outcomes as there are fewer medicines available, and
- **HER-2 positive** breast cancers (ER negative, PR negative, HER-2 positive) grow fast but are treatable with targeted hormonal therapies.

Most of the studies included in this literature review did not comment on the molecular subtype of detected cancers but five studies did. Both RCTs, the Malmö trial and one other Norwegian study reported data by molecular subtype (Hofvind et al., 2019; Johnson et al., 2019; Hofvind et al., 2018; Pattacini et al., 2018). One other study (Houssami et al., 2019) reported on dominant receptors but not molecular subtype. Data from individual studies is presented in *Tables 11a* and *11b* (overleaf).

[17] Receptors include oestrogen and progesterone receptors (OR and ER respectively). A key protein involved in growth of cancer is the human epidermal growth factor receptor 2 (HER-2). Oestrogen, progesterone and HER-2 stimulate cell growth and are important contributors to the aggressiveness of cancers. There are a range of treatments for many hormone-sensitive cancers which positively affect prognosis. A triple negative cancer (ER-,PR-, HER-2-) does not respond to these treatments and may have a poorer prognosis.

[18] Ki67 is a biopsy-based laboratory test that uses the presence of Ki67 protein to determine the aggressiveness of a cancer. Large amounts of Ki67 indicate faster cell proliferation. Results over 20% indicate a more aggressive tumour but studies reported different cut-off ranges (eg, ≤20% or ≤30%).

Table 11a: Molecular subtype and/or receptor status for a range of imaging protocols

Study	Number of cancers	DBT+s2DM	FFDM+DBT	FFDM alone
DBT+s2DM compared to FFDM alone				
RCT				
Hofvind et al. (2019) To-Be-1 (main analysis)	Total: 182 DBT+s2DM: 95 FFDM: 87	Luminal A: 58.7% (46.7, 69.9) Luminal B HER-2-ve: 24% (14.9, 35.3) Luminal B HER-2+ve: 6.7% (2.2, 14.9) Triple negative: 6.7 % (2.2, 14.9) HER-2+: 10.1% (4.5, 19) OR+: 89.9% (81.0, 95.5) PR+: 77.2% (66.4, 85.9) Ki67≥30%: 29.6% (19.3, 41.6)	NA	Luminal A: 61.4% (49, 72.8, *p*=.43) Luminal B HER-2-ve: 25.7% (16.0, 37.6) Luminal B HER-2+ve: 10.0% (4.1, 19.5) Triple negative: 1.4% (0.04, 7.7) HER-2+: 11.3% (11.1, 13.8; *p*=.82) OR+: 97.2% (90.2, 99.7; *p*=.07) PR+: 87.3% (77.3, 94.0; *p*=.11) Ki67≥30%: 20.0% (11.1, 31.8; *p*=.20)
Prospective studies embedded in population-based screening				
Houssami et al. (2019) Maroondah	DBT+s2DM: 49 FFDM: 34	HER-2+: 14% HER-2-: 87% OR/PR+: 97% OR/PR-: 3%	NA	HER-2+: 0 HER-2-: 100% OR/PR+: 93% OR/PR-: 7%
Hofvind et al. (2018) Norway BSP	DBT+s2DM: 348 FFDM: 379	Luminal A: 4.2 per 1000 Luminal B HER-2-ve: 2.6 per 1000 Luminal B HER-2 +ve: 0.3 per 1000 HER-2 +ve: 0.03 per 1000 Triple negative: 0.3 per 1000	NA	Luminal A: 3.2 per 1000 (*p*=.001) Luminal B HER-2 -ve: 1.5 per 1000 (*p*=.05) Luminal B HER-2 +ve: 0.3 per 1000 HER-2 +ve: 0.1 per 1000 Triple negative: 0.2 per 1000
Retrospective analysis				
Hovda et al. (2019) Norway BSP	FFDM+FFDM: 39 FFDM+DBT or DBT+s2DM: 248	Luminal A: 6.06 per 1000 Luminal B HER-2+ve: 0.25 per 1000 Luminal B HER-2-ve: 1.25 per 1000 HER-2+ve: 0.07 per 1000 Triple negative: 0.47 per 1000	NA	Luminal A: 2.09 per 1000 Luminal B HER-2+ve: 0.29 per 1000 Luminal B HER-2-ve: 0.48 per 1000 HER-2+ve: 0.19 per 1000 Triple negative: 0.67 per 1000
FFDM+DBT to FFDM alone				
RCT				
Pattacini et al. (2018) Reggio Emilia	Total: 127 FFDM+DBT: 83 FFDM: 44	NA	PR>10%: 62 cancers ER≥10%: 75 cancers HER-2 -ve: 58 cancers Ki67 <20%: 50 cancers	PR>10%: 26 cancers (RR=2.30; 1.5, 3.77) ER≥10%: 40 cancers (RR=1.88; 1.28, 2.75) HER-2 -ve: 28 cancers (RR=2.07; 1.32, 3.25) Ki67 <20%: 27 cancers (RR=1.85; 1.16, 2.96)
Retrospective analysis				
Dang et al. (2019)	FFDM+DBT: 37 FFDM: 26	NA	PR+ve: 74% ER+ve: 93%	PR+ve: 83% (*p*=.46) ER+ve: 93% (*p*=.34)

Table 11b: Molecular subtype and/or receptor status DBT_{MLO} compared to FFDM alone

Study	Number of cancers	Detected on DBT_{MLO} only	Detected on FFDM
Prospective studies embedded in population-based screening			
Johnson et al. (2019) Malmö	Total: 118 cancers Additional invasive cancers only detected with DBT_{MLO}: 37 *NB no statistically significant differences reported; some cancers did not have data which accounts for the difference in percentage rates (usually only one or two cancers were missing)*	Luminal A: 51% Luminal B HER-2-ve: 35% Luminal B HER-2+ve: 0 Triple negative: 8% HER-2+ amplified: 3% OR+ >10%: 86% OR- ≤10%: 11% PR+ >10%: 68% PR- ≤10%: 30% Ki67≤20%: 57% Ki67>20%: 41%	Luminal A: 46% Luminal B HER-2-ve: 43% Luminal B HER-2+ve: 1% Triple negative: 5% HER-2+ amplified: 5% OR+ >10%: 90% OR- ≤10%: 10% PR+ >10%: 74% PR- ≤10%: 25% Ki67≤20%: 58% Ki67>20%: 43%

Mixed results are presented about whether the cancers detected by DBT differed in terms of molecular subtype to those detected with FFDM. For example, in the main analysis from the To-Be-1 RCT Hofvind et al. (2019) reported no significant difference between the molecular subtype of cancers detected with DBT+s2DM compared to FFDM alone or between any of the receptors or Ki67 percent. Johnson et al. (2019) reported data on the tumour characteristics and molecular subtypes of cancers detected in the Malmö screening trial, which used a different imaging protocol (DBT_{MLO} compared to FFDM). This was one of the few studies to look at cancer characteristics by additional cancers detected with DBT (compared to all cancers in the study dataset or a comparison of all cancers with DBT+s2DM or FFDM and all cancers detected with FFDM alone). No statistically significant differences Data in molecular subtype were identified in the Malmö trial data.

Other studies (including Hofvind et al.'s 2018 and Hovda et al.'s retrospective analysis of data from the Norway BreastScreen Program and baseline RCT data from Pattacini et al.'s trial) showed an increase in cancers with a molecular profile more favourable to good prognosis. Hofvind et al. (2018) reported differences in subtype in their comparative data of one DBT+s2DM round compared to two previous FFDM rounds noting more Luminal A and Luminal B (HER-2 negative) breast cancers (but no statistical testing was completed). Statistical testing was completed in Hovda et al.'s 2019 work (also completed in the Norway BreastScreen Program). They reported statistically significant increases in the number of Luminal A (an increase of 3.97 cancers per 1000 screens) and Luminal B HER-2 negative (an increase of 0.77 cancers per 1000 screens) cancers detected in paired screening data for women who had DBT after a previous FFDM round (whether FFDM+DBT or DBT+s2DM). Houssami et al. (2019) reported receptor status for DBT+s2DM and FFDM but did not complete any comparative analysis due to the small number of cancers. Although more HER-2 positive cancers were detected in the DBT arm of the trial, the authors make no comment on this.

The other RCT (Pattacini et al., 2018) reported on differences in receptor status (not specific molecular subtype) for cancers detected with FFDM+DBT compared to FFDM alone. In the DBT arm of the trial, detected cancers were more likely to have more favourable biomarkers (i.e., PR positive, ER positive, HER-2 negative, and low Ki67), and that, when coupled with data on tumour size and grade was suggestive of earlier detection of clinically important cancers. Note that the results in *Table 11a* refer only to significant RR data.

There is some evidence to suggest that DBT may be detecting more cancers with a molecular subtype that has a more favourable prognosis and which are better depicted on DBT, or it may be detecting cancers that are less likely to progress aggressively. Further Lång (2019) commented that the possible increase in Luminal A cancers could be due to the presentation of these cancers as architectural distortions (AD), which is much more conspicuous on DBT imaging compared to FFDM images. Lack of consistent evidence may reflect the small number of cancers covered in some studies or differences in the groups of cancers that are compared to each other in the primary studies. For example, Johnson et al. (2019) compared invasive cancers only detected with DBT_{MLO} to all cancers detected with digital mammography. Other studies did not compare only additional cancers detected with DBT, rather they compared data from the DBT arm to the FFDM arm. This may underestimate any differences between additional cancers detected with DBT. Also, different studies reported information in different ways (by molecular subtype or by receptor status only) and many studies only include a small number of cancers, giving limited ability to perform adequate statistical comparison (i.e., Maroondah data was descriptive and the Norway BreastScreen Program study did no statistical testing).

There is not yet sufficient evidence to determine if additional cancers detected with DBT alone are more aggressive 'killing' cancers or if they are comparable to the types of cancers detected with FFDM (but more are detected). More research in this area is needed to determine if the additional cancers detected are clinically significant from a molecular perspective.

3.2.3. Cancer staging

The assessment of newly diagnosed breast cancer is essential to obtain an estimate of staging, which is integral in prognosis development. Staging is used to describe the specific characteristics of the cancer. It involves determining the extent of disease in the affected breast and in the contralateral breast, evaluating regional lymph nodes and identifying other sites of disease if the cancer has metastasised. Staging is also used to assist in treatment planning and informs follow-up surveillance. The TNM classification along with other measures are used to determine a patient's:

- overall cancer stage (stage 0 – stage IV)
- primary tumour size (T)
- whether regional nodes are involved (N), and
- distant metastasis (M).

Most of the studies evaluated in this literature review did not comment directly on staging using the TNM framework, rather data was reported by the individual components of this framework. For this reason, this section also presents data by staging framework component. Staging data is provided where information is presented (Bernardi et al., 2019; Caumo et al., 2018). No studies evaluated in this literature review reported on metastasis.

Only three of the population-based embedded studies reported on cancer stage generally (as well as reporting on the separate components); all other studies reported on the components of staging (eg, histological grade, tumour size and nodal status). Two of these studies related to DBT+s2DM compared to FFDM (Trento and Verona, Bernardi et al., 2019 and Caumo et al., 2018) and one compared FFDM+DBT to FFDM alone (Reggio Emilia RCT, Pattacini et al., 2018).

Bernardi et al. (2019) reported that DBT+s2DM detected more stage 1 and stage 2 cancers than FFDM alone (grade 1 cancers RR=1.72; 95%CI: 1.38, 2.13; grade 2 cancers RR=1.47; 95%CI: 1.04, 2.08). The authors concluded that use of DBT resulted in increased detection for all stages of cancer except for DCIS (see previous section) and thought that staging evidence suggested that DBT does not contribute to overdiagnosis. Likewise, data from the Verona trial reported that less stage 0 (12.3 percent of cancers detected with DBT+s2DM compared to 23.1 percent detected with FFDM alone in previous screening rounds) and more stage 1a tumours were detected with DBT+s2DM (63 percent compared to 50 percent), stage 1b (9.7 percent compared to 2.6 percent); however, none of these results reached statistical significance but the authors concluded that more invasive cancers of clinical interest were detected at an earlier stage and that DBT may confer a mortality benefit.

For FFDM+DBT, Pattacini et al. (2018) reported on cancer grade, reporting that DCIS and more stage I cancers were detected in the DBT arm (2.80; 95%CI: 1.01, 7.65 and RR=1.89; 95%CI: 1.20, 2.99 respectively) but no difference was reported for Grade II and above. The wide confidence intervals in the RR calculation are likely to reflect the small number of cancers detected in this study.

3.2.4. Histological grade (invasive cancers only)

Breast cancer grading is conducted histologically after a successful biopsy is completed and depends on how the tumour cells differ from healthy cells. Grade 1 breast cancer cells look small and uniform like healthy cells and are usually slow growing. Grade 3 breast cancer cells appear abnormal, usually due to a much faster rate of growth. *Table 12* includes results from all studies.

DBT+s2DM compared to FFDM alone

RCT

Results from the To-Be-1 RCT (Hofvind et al., 2019) reported that use of DBT+s2DM did not clearly indicate whether DBT+s2DM detected more Grade 1 cancers compared to those detected with FFDM alone.

Prospective studies embedded in screening programs

A number of prospective studies embedded in population-based screening programs reported results by histological grade for DBT+s2DM compared to FFDM alone; however, reported results were inconsistent with those reported in the To-Be-1 RCT. Other studies (including results from the Maroondah, Trento, Cordoba and Norway BreastScreen program studies) generally reported an increase in detection of Grade 1 cancers with DBT. Often, the increase in grade 1 cancers was about double with DBT+s2DM (see *Table 12* for further specific detail). Authors cautioned about these results based on small numbers of cancers.

FFDM+DBT compared to FFDM alone

Prospective studies embedded in screening programs

When FFDM+DBT was compared to FFDM alone, non-significant increases in the detection of Grade 1 cancers were reported (Pattacini et al., 2018) and no differences by imaging protocol were reported in the OTS data.

DBT_{MLO} compared to FFDM alone

Prospective studies embedded in screening programs

No difference in cancer grade was detected in the Malmö trial (Zackrisson et al., 2018).

Table 12: Histological grade for a range of imaging protocols

Study	DBT+s2DM	DBT_{MLO}	FFDM+DBT	FFDM alone
DBT+s2DM compared to FFDM alone				
RCT				
Hofvind et al. (2019) To-Be-1 RCT (main analysis)	Grade 1: 28.9% (19.1, 40.5) Grade 2: 50% (38.3, 61.7) Grade 3: 21.1% (12.5, 31.9) No data: 5%	NA	NA	Grade 1:34.8% (23.7, 47.2) Grade 2: 50.7% (38.4, 63) Grade 3: 14.5% (7.2, 52) No data: 2.8%
Prospective studies embedded in population-based screening				
Bernardi et al. (2019) Trento 205 cancers	Grade 1: 29.43% Grade 2: 51.12% Grade 3: 19.45%	NA	NA	Grade 1: 23.41% Grade 2: 57.08% Grade 3: 19.51%
Houssami et al. (2019) Maroondah	Grade 1: 45% Grade 2: 48% Grade 3: 8%	NA	NA	Grade 1: 20% Grade 2: 53% Grade 3: 27%
Hofvind et al. (2018) Cancers per 1000	Grade 1: 3.3 per 1000 Grade 2: 3.5 per 1000 Grade 3: 0.8 per 1000 No data: 2 cancers	NA	NA	Grade 1: 1.4 per 1000 (p<.001) Grade 2: 3.0 per 1000 (p=.233) Grade 3: 0.9 per 1000 (p=.376) No data: one cancer
Romero Martín et al. (2018) Cordoba	*IDC* Grade 1: 33 cancers Grade 2: 24 cancers Grade 3: 10 cancers *ILC* Grade 1: 2 cancers Grade 2: 2 cancers *DCIS:* no difference in grade for in situ cancers	NA	NA	*IDC* Grade 1: 23 cancers (p=.001) Grade 2: 23 cancers (p=.552) Grade 3: 9 cancers *ILC* Grade 1: 2 cancers Grade 2: 0 cancers
Retrospective analysis				
Hovda et al. (2019) Norway BSP	Grade 1: 3.17 per 1000 Grade 2: 3.74 per 1000 Grade 3: 1.28 per 1000	NA	NA	Grade 1: 0.95 per 1000 (p=.001) Grade 2: 1.24 per 1000 (p=.001) Grade 3: 1.52 per 1000
FFDM+DBT compared to FFDM alone				
RCT				
Pattacini et al. (2018) Reggio Emilia	NA	NA	Grade 1: 12 cancers Grade 2: 61 cancers Grade 3: 8 cancers No data: 2 cancers	Grade 1: 4 cancers (RR3.00; 0.97, 9.3) Grade 2: 30 cancers (RR=2.03, 1.32, 3.15)

Study	DBT+s2DM	DBT_{MLO}	FFDM+DBT	FFDM alone
				Grade 3: 9 cancers (RR=0.89 0.34, 2.30) No data: 1 cancer
Prospective studies embedded in population-based screening				
Skaane et al. (2018) OTS	NA	NA	Grade 1: 35.4% Grade 2: 46.0% Grade 3: 18.5% No data: 2 cancers	Grade 1: 35.8% Grade 2: 45.9% Grade 3: 18.0% No data: NA
Retrospective studies				
Bahl et al. (2019)	NA	NA	*Invasive cancers* Grade 1: 20.0% Grade 2: 55.6% Grade 3: 22.9% No data: 2 cancers	*Invasive cancers* Grade 1: 19.1% (*p*=.93) Grade 2: 52.2% Grade 3: 23.6% No data: NA
Dang et al. (2019)	NA	NA	Grade 1: 37% Grade 2/3: 63%	Grade 1: 28% Grade 2/3: 72% (*p*=.52)
DBT_{MLO} compared to FFDM alone				
Prospective studies embedded in population-based screening				
Johnson et al. (2019) Malmö	NA	Grade 1: 41% Grade 2: 46% Grade 3: 11% No data: 3%	Grade 1: 40% Grade 2: 43% Grade 3: 16% No data: 1%	NA
Zackrisson et al. (2018) Malmö	NA	Grade 1: 40% Grade 2: 43% Grade 3: 15% No data: 2%	Grade 1: 40% Grade 2: 43% Grade 3: 16% No data: 1%	NA

3.2.5. Tumour size (mm)

BSA defines a small tumour as less than 15mm in diameter. Studies evaluated as part of this literature review used different bands to describe the size of detected invasive cancers, which made it difficult to compare results; however, mixed results were presented between the RCT and the other studies embedded in population-based screening programs. Individual study results are reported in *Table 13* (overleaf).

DBT+s2DM compared to FFDM alone

Most studies suggested that smaller cancers are detected with DBT+s2DM. The clinical significance of the detection of smaller cancers is uncertain (and needs to be considered in light of other tumour characteristics).

RCT

For DBT+s2DM compared to FFDM, To-Be-1 RCT data reported a similar proportion of cancers detected ≤20mm but fewer very small cancers (≤10mm) were detected in the DBT arm (Hofvind et al., 2019) (although this result was not significant). In the To-Be-1 RCT, cancers detected with DBT had larger mean and median diameters; however, the results for tumours ≤20mm were similar between the two imaging protocols.

Prospective studies embedded in screening programs

Romero Martín et al. (2018) and Hofvind et al. (2018) reported a statistically significant increases in the detection of cancers ≤10mm with DBT+s2DM compared to FFDM alone. Results from other prospective studies also suggested that DBT+s2DM detected proportionally more very small cancers (≤5mm) compared to FFDM alone (Bernardi et al., 2019; Houssami et al., 2019). Data from Maroondah suggested that DBT+s2DM detected more very small cancers (≤5mm) but both study arms had similar results for detecting small cancers (using the BSA definition of less than 15mm): 61 percent of cancers detected with DBT+s2DM were under 15mm compared to 60 percent with FFDM imaging (Houssami et al., 2019). This result was similar for cancers detected in the Trento screening program (although a direct comparison is difficult because the results are stratified by different size bands) (Bernardi et al., 2019). Authors generally commented that, proportionally, more smaller cancers were detected with DBT or that DBT-detected cancers had a slightly smaller mean diameter, although not all studies reported statistically significant results. Median diameter was also slightly smaller (when this measure was reported).

No mean size was provided for the Trento data, but was the only data provided in the Verona study (Caumo et al., 2018), which reported that the mean diameter of tumours detected with DBT+s2DM was 3.5mm smaller than those cancers detected with FFDM. This difference was much larger than the difference in mean presented in the Maroondah data, although the direction of effect is the same (i.e., more smaller cancers detected with DBT). The strongest evidence that DBT+s2DM detects more smaller cancers comes from Hofvind et al.'s 2018 analysis of Norway BreastScreen Program data reported a statistically significant increase in the number of cancers under ≤10mm and for cancers between 10 and 20mm in diameter.

Retrospective analysis

Hovda et al. (2019) reported an increase in the detection of very small cancers with similar results presented in their retrospective analysis as that presented in the prospective studies.

FFDM+DBT compared to FFDM alone

Results from studies comparing FFDM+DBT to FFDM alone also reported that the DBT arm resulted in detection of smaller cancers (i.e., both those ≤10mm and ≤20mm) (Pattacini et al., 2018; Rose & Shisler, 2018).

RCT

The data presented from the Reggio Emilia trial included 69/83 invasive cancers detected with FFDM+DBT and 39/44 invasive cancers detected with FFDM alone but none of the DCIS). Other cancer characteristics reported on included DCIS (eg, grade, stage, etc.). Data relating to invasive cancer only showed an increase in the proportion of cancers detected with FFDM+DBT were ≤10mm compared to FFDM alone. In this study, women with a family history and those having a prevalent screen were excluded, which the authors noted could mean fewer large cancers were detected during the study (and this could have influenced the non-significant results for detection of cancers ≥20mm).

Prospective studies embedded in screening programs

No significant differences in tumour size were reported in the OTS trial (Skaane et al., 2018).

DBT_{MLO} compared to FFDM alone

Prospective studies embedded in screening programs

Results from the Malmö trial were equivocal with similar numbers of small cancers detected in both the DBT_{MLO} arm and the FFDM arm (Zackrisson et al., 2018). The authors compared the difference in mean diameter size for cancers only detected with DBT and reported that these cancers were quite a bit smaller: 76 percent of cancers only detected with DBT were under 15mm compared to 66 percent of all cancers detected with FFDM.

Table 13: Tumor size (mm) for invasive cancers for a range of imaging protocols

Study	DBT+s2DM	DBT_{MLO}	FFDM+DBT	FFDM alone
DBT+s2DM compared to FFDM alone				
RCT				
Hofvind et al. (2019) To-Be-1 RCT (main analysis)	≤10mm: 14.3% 10-20mm: 61.4% ≥20mm: 24.3% No data: 12.5% Mean: 16mm (SD 8.4) Median: 14.9 (11, 18)	NA	NA	≤10mm: 30.0% (*p*=.09) 10-20mm: 48.3% ≥20mm: 21.7% No data: 15.5% Mean: 14.5 (SD 8.8; *p*=.33) Median: 14.0 (8.5, 18.8)
Prospective studies embedded in population-based screening				
Bernardi et al. (2019) Trento	≤5mm: 7.48% 5.1-10mm: 29.92% 10.1-20mm: 36.16% 21-49mm: 12.72% ≥50mm: .997%	NA	NA	≤5mm: 4.88% 5.1-10mm: 26.83% 10.1-20mm: 38.05% 21-49mm: 12.19% ≥50mm: 1.95%
Houssami et al. (2019) Maroondah	≤5mm: 15% 5.1-10mm: 13% 10.1-15mm: 33% 15.1-20mm: 20% ≥20mm: 20% Mean: 16.4mm (SD 12.8mm)	NA	NA	≤5mm: 13% 5.1-10mm: 27% 10.1-15mm: 20% 15.1-20mm: 17% ≥20mm: 23% Mean: 16.8 mm (SD 12.3mm)
Caumo et al. (2018) Verona	Mean: 15.1mm	NA	NA	Mean: 18.6mm (*p*=.052)
Hofvind et al. (2018) Norway BSP	≤10mm: 3.2 per 1000 10-20mm: 3.3 per 1000 ≥20mm: 1.0 per 1000 No data: 4 cases	NA	NA	≤10mm: 1.8 per 1000 (*p*<.001) 10-20mm: 2.5 per 1000 (*p*=.03) ≥20mm: 0.9 per 1000 (*p*=.747) No data: nine cancers
Romero Martín et al. (2018) Cordoba	≤10mm: 25 cancers 11-15mm: 14 16-19mm: 14 ≥20mm: 39	NA	NA	≤10mm: 19 cancers (*p*=.021) 11-15mm: 12 16-19mm: 12 ≥20mm: 33
Retrospective analysis				
Hovda et al. (2019)	<10mm: 1.52 per 1000 10-19mm: 1.05 per 1000 20-29mm: 0.76 per 1000 ≥30mm: 0.38 per 1000	NA	NA	<10mm: 3.33 per 1000 (*p*=.01) 10-19mm: 3.74 per 1000 (*p*=.001)

Study	DBT+s2DM	DBT_{MLO}	FFDM+DBT	FFDM alone
	Mean: 15.2mm Median: 12mm			20-29mm: 0.74 per 1000 ≥30mm:0.44 per 1000 Mean: 13.3mm Median: 11mm
FFDM+DBT compared to FFDM alone				
RCT				
Pattacini et al. (2018) Reggio Emilia	NA	NA	≤10mm: 31 cancers 10-20mm: 31 cancers ≥20mm: 7 cancers	≤10mm: 16 cancers (RR=1.94; 95%CI: 1.06, 3.54) 10-20mm: 14 cancers (RR=2.22; 95%CI: 1.18, 4.16) ≥20mm: 8 cancers (RR=0.88; 0.32, 2.41) No data: 1
Prospective studies embedded in population-based screening				
Skaane et al. (2018) OTS	NA	NA	≤10mm: 9.3% 11-20mm: 41.9% >20mm: 48.8%	≤10mm: 17.2%, *p*=.138 11-20mm: 45.2% >20mm: 37.6%
Retrospective analysis				
Bahl et a. (2019)	NA	NA	Mean: 12.0mm (±9.7)	Mean: 12.8mm (±8.2)
Rose & Shisler (2018)	NA	NA	Mean: 14mm (6, 26) Median: 15mm	Mean: 17mm (4, 60) Median: 19mm
DBT_{MLO} compared to FFDM alone				
Prospective studies embedded in population-based screening				
Johnson et al. (2018) Malmö	NA	≤10mm: 41% 10-20mm: 43% 20-50mm: 14% ≥50mm: 0%	≤10mm: 32%*p*=.88 10-20mm: 52% 20-50mm: 14% ≥50mm: 1%	NA
Zackrisson et al. (2018) Malmö	NA	≤10mm: 37% 11-15mm: 33% 16-19mm: 11% ≥20mm: 18% Mean: 14mm Median: 12mm (1,78)	≤10mm: 35% 11-15mm: 31% 16-19mm: 10% ≥20mm: 23% Mean: 15mm Median: 12mm (1, 78)	NA

Taken together, these study results suggest that imaging with DBT may result in the detection of slightly smaller cancers, most of which are likely to be invasive (see previous section on *cancer type*). This suggests that use of DBT may not increase overdiagnosis due to finding more DCIS.

3.2.6. Lymph node involvement

Early stage breast cancers have not yet spread to surrounding lymph nodes (i.e., node negative disease). Detection of node negative disease is clinically important because detecting these cancers at screening may mean women are more likely to have a better prognosis as the cancer has not yet spread. Most of the key studies (including the To-Be-1 RCT and the studies embedded in screening programs) reported on the proportion of cancers that were node-negative or node-positive. Most reported no or only small differences in the proportion of node negative and node positive disease, regardless of the imaging protocol used (i.e., DBT+s2DM, FFDM+DBT or DBT_{MLO}).

If there was a difference, it usually favoured the DBT imaging arm (i.e., DBT detected more node-negative disease). For example, quite a large increase in node-negative disease when DBT was used was reported in Hovda et al.'s 2019 analysis (comparing women who had two consecutive screens with FFDM compared to women whose first screen was FFDM and the subsequent screen was either FFDM+DBT or DBT+s2DM). Results suggesting slightly more detection of node-positive disease were also reported in the Verona Screening Program evaluation. Results are presented in *Table 14* (below).

Table 14: Nodal status for a range of imaging protocols

Study	DBT+s2DM	DBT_{MLO}	FFDM+DBT	FFDM alone
DBT+s2DM compared to FFDM alone				
RCT				
Hofvind et al. (2019) To-Be-1 RCT (main analysis)	Positive: 17.7% (10.0, 27.9)	NA	NA	Positive: 25.7% (16.0, 37.6; *p*=.24)
Prospective studies embedded in population-based screening				
Bernardi et al. (2019) Trento	Positive: 16.46% Negative: 74.06% Unknown: 9.48%	NA	NA	Positive: 17.07% Negative: 71.22% Unknown: 11.71%
Houssami et al. (2019) Maroondah	Positive: 15% Negative: 85%	NA	NA	Positive: 20% Negative: 80%
Caumo et al. (2018) Verona	Positive: 26.9%	NA	NA	Positive: 11.5% (*p*=.032)
Hofvind et al. (2018) Norway BSP	Positive: 1 per 1000	NA	NA	Positive: 0.7 per 1000, (p=.341)
Retrospective studies				
Hovda et al. (2019) Norway BSP	Positive: 1.28 per 1000 Negative: 6.80 per 1000 (*p*=.001)	NA	NA	Positive: 0.57 per 1000 Negative: 3.14 per 1000
FFDM+DBT compared to FFDM alone				
Prospective studies embedded in population-based screening				
Skaane et al. (2018) OTS	NA	NA	Positive: 15.6% Negative: 84.4% Unknown: five cancers	Positive: 21.4% Negative: 78.6% Unknown: three cancers
Retrospective studies				
Bahl et al. (2019)	NA	NA	Positive: 10.2%	Positive: 16.6% (*p*=.054)
Dang et al. (2019)	NA	NA	Positive: 15% Negative: 85%	Positive: 17% Negative: 83% (*p*=.87)
DBT_{MLO} compared to FFDM alone				
Prospective studies embedded in population-based screening				
Zackrisson et al. (2018) Malmö	NA	Positive: 22% Negative: 69% NA: 8%	Positive: 22% Negative: 67% NA: 5%	NA

3.3. Radiological presentation

Suspicious mammographic findings may include distortion of normal breast tissue without a corresponding mass, architectural distortion (AD) a difference in breast density between the breasts (a global or focal asymmetry), specific patterns of microcalcifications or masses. Some primary mammographic features are clearly seen on FFDM (such as microcalcifications and some masses), whereas other presentations (like AD or asymmetries) can be subtle on FFDM and are much more difficult to detect. Using methods that improve visibility and clarity of primary mammographic findings (especially subtle findings like AD or focal asymmetry) has a positive impact on readers' ability to detect areas suspicious for cancer and can improve diagnostic accuracy and reader performance. We want to know, based on current evidence, whether DBT provides superior conspicuity of cancers by radiological presentation.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

Allen + Clarke's literature review of the role of tomosynthesis in the assessment of screen-detected abnormalities found that compared to FFDM, DBT improved the conspicuity of cancers presenting as AD and increased the overall detection of subtle AD (whether related to a malignant or a benign structure). Evidence suggested that AD presentation visible on DBT but occult on FFDM or ultrasound should be treated as suspicious for cancer given the high PPV_3. All studies (almost all of which were in mixed study populations) reported improved conspicuity of asymmetry with DBT. The shape, texture and appearance of tissue in and around a mass is indicative of whether a mass is benign or suspicious for malignancy. Imaging with DBT often resulted in reclassification of primary mammographic finding from a focal asymmetry to a mass. Improved lesion conspicuity with DBT reduced tissue overlap making it easier to determine mass margin by removing 'noise' and making it easier to determine between masses suspicious for malignancy and those that were benign. There was sufficient evidence that DBT provides superior performance in terms of improving readers' view of mass margins/soft tissue lesions. Bearing in mind that DCIS is easily seen on FFDM, overall, studies reported mixed results on whether DBT (either alone or as FFDM+DBT) had equivalent or inferior performance in terms of detecting microcalcifications compared to FFDM. There was some consensus that DBT alone may not be sufficient for the detection of cancers presenting with microcalcification as the primary mammographic finding. There was consensus that image quality is now equivalent to or in some cases better than FFDM for microcalcifications. Viewing DBT as a slab (rather than a slice) showed some promise in improving DBT's performance but further evidence of success was required.

Updated findings

Literature published since 31 December 2017 reported on some of the radiographical presentations of cancers detected. There was not much content on this topic (usually available only in tables and with little author comment). The results presented were consistent with previous findings. Because there is limited information about radiographical presentation, results are not segmented and presented by study type or in detailed data tables. Information is presented by radiological finding only. No systematic reviews of studies published since December 2017 have considered radiological presentation. Included studies are listed below.

Literature review

One literature review: Chong et al., 2019

Prospective studies embedded in population screening programs

Four studies (five papers): Skaane et al., 2019; Caumo et al., 2018; Romero Martín et al., 2018; Skaane et al., 2018; Zackrisson et al., 2018

Observational studies

Five studies with a retrospective design: Bahl et al., 2019, Choi et al., 2019; Wasan et al., 2019; Lai et al., 2018; Wahab et al., 2018

Summary of key findings published between 1 January 2018 and 31 December 2019

Only a small number of studies (n=9) reported information on radiological presentation by different imaging protocol. Most confirmed previously known information: use of DBT detects more cancers presenting as AD than FFDM, masses are better defined in a DBT image and therefore DBT detects more cancers presenting as spiculated and circumscribed masses, and that FFDM depicts microcalcifications very well but s2DM is not inferior.

3.3.1. Architectural distortion

Use of DBT results in increased detection of cancers presenting as AD due to improved conspicuity. This finding was consistent across a number of studies that reported on this presentation; however, we note that the overall number of cancers with and AD presentation is usually small.

In the Verona screening program evaluation (methodology described in *section 3.1.1*), Caumo et al. (2018) reported that they detected more cancers presenting as AD with DBT+s2DM (8.4 percent) compared to the previous screening rounds (5.1 percent; p=.40); however, this increase in detection was not significant. Data from the Malmö trial also showed an increase in detection but only three cancers were involved, two of which were detected with DBT and one with FFDM (Zackrisson et al., 2018). Skaane et al. (2019) also reported more cancers presenting as AD and fewer false positives with the use of DBT (either FFDM+DBT or DBT+s2DM compared to FFDM alone):

- FFDM alone: nine true positives (5.9 percent); 158 false positives (11.4 percent)
- FFDM+DBT: 31 true positives (15.7 percent); 153 false positives (12.7 percent), and
- DBT+s2DM: 28 true positives (14.4 percent); 162 false positives (14.6 percent).

Bahl et al. (2019) reported statistically significant results showing an increase in cancers presenting as AD with DBT (1.5 percent) compared to FFDM (1.4 percent, p=.04).

Romero Martín et al. (2018) also reported that all cancers presenting as AD were detected with DBT by both readers whereas only 40 percent of cancers presenting as AD were detected with FFDM, which indicated that cancers presenting as AD are easier to see on DBT. The authors also noted that AD is also often associated with benign lesions but that no difference in PPV_3 was detected between DBT and FFDM reading arms. Chong et al. (2019) also cited a small, retrospective study on AD only detected with DBT. DBT-only AD had a much lower PPV_3 (10.2

percent) compared to AD detected with FFDM (43.4 percent, $p<.001$), suggesting that more benign lesions presenting as AD are detected with DBT. Therefore, if an AD is only seen on DBT imaging, it may represent an additional assessment burden for a final benign outcome that could be avoided by using both DBT and a two-dimensional image (such as s2DM or DM_{CC} or DM_{MLO} depending on the DBT projection in which the AD is detected).

Skaane et al. (2018) also provided some final summary data from the OTS and reported that 14 of the cancers only detected with FFDM+DBT presented as AD (27 percent).

3.3.2. Asymmetric density

Only a few studies commented on cancers presenting as focal asymmetries and the number of cancers presenting as asymmetric density was extremely small. This did not provide much useful information (for example, Romero Martín et al. reported that three cancers presented as asymmetries, and two were detected with FFDM compared on one detected with DBT). Skaane et al. (2019) reported that, with the use of DBT, false positives associated with asymmetric density decreased (from 406 false positives with FFDM compared to 207 with FFDM+DBT or 168 false positives with DBT+s2DM). This may represent a decrease in recalls and investigations for these malignancy presentations when DBT is used.

3.3.3. Masses

DBT provides improved visualisation and conspicuity of masses, and researchers indicated that more cancers presenting as spiculated and circumscribed masses were detected with DBT compared to FFDM alone. For example, Romero Martín et al. (2018) reported that 100 percent of cancers presenting as mass were detected with DBT compared to 92.7 percent with FFDM. More masses were detected in Bahl et al.'s retrospective analysis: 5.6 percent of cancers detected with DBT presented as a mass compared to 1.4 percent of FFDM-detected cancers. Johnson et al. (2019) reported data from the Malmö trial but noted that there was not much difference between the two imaging protocols, although DBT was likely to result in more spiculated masses, which are often associated with ILC (which is less conspicuous on FFDM). Skaane et al. (2018) also provided final summary data from the OTS and reported that 27 of the cancers only detected with FFDM+DBT presented as spiculated masses (52 percent).

Wasan et al. (2019) assessed whether DBT can accurately predict if circumscribed masses are benign based on margin sharpness. Better delineation between circumscribed masses with a benign final outcome and those that are malignant is an important way to reduce the false positive recall rates in screening programs. The authors reviewed 122 lesions (drawn from the NHS UK screening population) and reported that, when FFDM was used, only 9.3 percent of masses were defined as having a margin visibility of greater than 50 percent (indicating better margin sharpness). When DBT was used, 50.5 percent of lesion had a margin visibility of greater than 50 percent. Margin visibility was better seen when DBT was used but the authors also noted that all radiological features need to be considered before deciding whether to recall someone for a mass.

3.3.4. Microcalcifications

FFDM is an excellent imaging tool for detecting microcalcifications. One of the challenges associated with implementing DBT in the screening environment is that the slice/slabbing techniques used may not provide superior visibility of microcalcifications compared to that seen

with FFDM; however, previous research suggested that DBT+s2DM is not inferior to this (see *Allen + Clarke*'s previous literature review). Research published since 31 December 2017 confirms this finding. Quick summary results include data from a detailed assessment of tumour characteristics from cancer detected in the Malmö trial (Johnson et al., 2019), which showed that DBT_{MLO} did not detect more cancers presenting as microcalcifications compared to FFDM because these are conspicuous on FFDM. Romero Martín et al. (2018) also observed no difference in cancers presenting as microcalcifications (double reading of both FFDM and DBT detected 80.9 percent of cancers presenting with calcification). Bahl et al. (2019) showed fewer cancers presenting as microcalcifications when DBT was used (28.3 percent compared to 36 percent with FFDM).

Chong et al. (2019) completed a literature on clinical practice use of DBT. They reported on two retrospective observer studies: Choi et al. (2019) and Lai et al. (2018) (also identified in our search). Both of these studies reported few differences between the ability of readers to detect cancers presenting as microcalcifications. In Lai et al.'s study, four readers compared DBT+s2DM to FFDM for the detection of microcalcifications, using 72 consecutive screening mammograms recalled because of a microcalcification presentation (18 of which were cancers) and 20 FFDM mammograms. Similar observer performance was reported for microcalcification assessment with DBT+s2DM and FFDM alone (DBT+s2DM kappa value: 0.63, $p<.001$; FFDM kappa value: 0.66, $p<.001$). Similar combined reader sensitivity results for cancers were reported for both DBT+s2DM (94 percent) and FFDM (92 percent). Combined reader specificity was also similar: for cancers presenting as microcalcifications DBT+s2DM was 95 percent, for FFDM it was 98%. Choi et al. (2019) reviewed 198 cancers each with DM, s2DM and DBT images (i.e., investigative images). They reported that, in terms of diagnostic performance, there was little difference in AUROC measures. Wahab et al. (2018) also completed a two-reader retrospective review to determine if s2DM was equivalent to FFDM for the detection of calcifications in terms of BIRADS assessment. They reported that there was moderate to substantial agreement between the two imaging methods, concluding that s2DM can be used instead of FFDM with no decrease detection of cancers presenting as microcalcifications. Together, these studies build on the advice in *Allen + Clarke*'s literature on the role of DBT in the assessment of lesions, suggesting that DBT+s2DM is not inferior to FFDM alone. Chong et al. (2019) reported on a number of clinical factors needed to ensure that DBT+s2DM can be used effectively in the screening environment, including:

- Assessing images during acquisition to ensure motion is minimised
- Ensuring the processing algorithm is correct, acknowledging the higher contrast resolution with s2DM algorithms may increase conspicuity of microcalcifications, and structural or quantum noise is minimised (and the reconstructed image is true), and
- Careful scrolling by radiologists through slabs and stacks and through both MLO and craniocaudal projections (i.e., reader learning is important).

3.4. 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.

An 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 but definitions differed across studies). A negative initial or final assessment is a false negative if a breast cancer is subsequently diagnosed within a specified follow-up time. 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 both increase cancer detection and appropriately reduce recalls (especially false positive recalls, i.e., unnecessary recalls from screening for women who do not have breast cancer).

Section 3.4 describes overall recall rate, false positive recall rate and relative specificity.

3.4.1. Overall recall rates

The overall recall rate is the percentage of women asked to return for further tests after an abnormality is detected on a screening mammogram (i.e., any recall resulting in true or false positive findings). A lower frequency of recall is beneficial for women if the proportion of screen-detected breast cancer remains stable or increases. There is also a 'baseline' recall rate, below which concerns about program safety might be raised as not enough screening examinations need to be assessed. Reader protocols (single or double, with and without consensus) have the potential to impact recall rates, with lower recall rates generally observed by programs that employ double reading (i.e., the European and Australian programs) compared to those that employ single reading (i.e., the American screening programs).

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

The literature was not settled about the association between use of DBT as a screening test and overall recall rates. Two main complexities existed:

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 in double-reading programs but others reported reductions, and different studies used different interpretation or arbitration protocols which may influence recall rates, and
2. The limited room for improvement if recall rates are already appropriately low.

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). Both Coop et al. (2016) and Hodgson et al. (2016) 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 were 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.

Some prospective trials reported increased recall with double reading (either by two radiologists or through an arbitration process) but reduced false positive rate. Others reported no difference or a decrease. 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 without compromising the quality of the program). For example, 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 percent 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 percent for FFDM+DBT compared with 2.1 percent 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. Post-arbitration, this translated to FFDM+DBT having 6.2 more recalls per 1000 screens than FFDM alone.

Most of the studies investigating recall rates had short timeframes (≤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 was 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.

Information from the smaller retrospective studies (most of which used single reading strategies) reported that recall rate was reduced with FFDM+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 DBT images and availability of prior DBT images may also further decrease recall rates.

GRADE assessment: overall recall rate: FFDM+DBT compared to FFDM alone

Participants Studies	Quality of evidence	Overall results
572,555 participants 15 studies	⨁⨁ Low	Pooled analysis and data from prospective fully paired trials and retrospective analysis shows inconsistent results.

GRADE assessment: overall recall rate: DBT+s2DM compared to FFDM+DBT or FFDM alone

Participants Studies	Quality of evidence	Overall results
148,814 participants 4 studies	⨁⨁ Low	Pooled analysis and data from prospective fully paired trials and retrospective analysis shows inconsistent results.

GRADE assessment: overall recall rate: DBT_{MLO} compared to other imaging

Participants Studies	Quality of evidence	Overall results
7,681 2 studies	⊕ Very low	No systematic review or pooled analysis was available.

Updated findings

Literature published since 31 December 2017 reported on overall recall rates. As BSA standard reader protocol is independent double reading, we focused most attention on studies with similar policy settings (rather than those with annual screening and single reading). 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. Detailed methodologies for each of these studies is provided in *section 3.1.1.* Included studies are listed below. *Tables 15a-c* includes a summary of all study results.

Systematic reviews

Two systematic reviews: Marinovich et al., 2018, Phi et al., 2018

Randomized controlled trials

One RCT (two papers): Aase et al., 2019; Hofvind et al., 2019

One randomized study embedded in a population screening program: Pattacini et al., 2018

Prospective trials embedded in population screening programs

Eight studies: Bernardi et al., 2019; Houssami et al., 2019; Skaane et al., 2019; Caumo et al., 2018; Hofvind et al., 2018; Romero Martín et al., 2018; Skaane et al., 2018, Zackrisson et al. (2018)

Observational studies

One study with a prospective design: Miglioretti et al., 2019

Six studies with a retrospective design: Bahl et al., 2019; Conant et al., 2019; Honig et al., 2019; Hovda et al., 2019; Rose & Shisler, 2018; Upadhyay et al., 2018

Summary of key findings published between 1 January 2018 and 31 December 2019

The effect of DBT on recall rate varies as does its importance: a decrease in recall rate with DBT may be appropriate in a program which has high recall rates overall but may be less important where rates are already appropriately low.

Two recent systematic reviews indicated mixed results between paired and unpaired studies. In studies where recall rate is quite low already (often those using double reading plus some form of arbitration/consensus), recall rates with DBT were usually a little higher with DBT modalities. The pooled analysis suggested that overall there was a 1.8 percent decrease in recall with FFDM+DBT compared to FFDM alone. In studies where single reading was the norm, recall rates may decline more with the use of DBT. Many studies also observed a learning curve effect, with recall rates for DBT declining in time as readers learnt some of the subtle differences in presentation of cancers on DBT (for example, presentation as architectural distortion). That said, observed recall rates may still have been higher with DBT even after improvements.

Of particular relevance to the BSA were the findings of the Maroondah pilot trial, which reported an increase in recall for DBT+s2DM compared to FFDM alone; however, the overall increase in recall was not extremely large (4.2 recalls per 1000 women screened with DBT+s2DM compared to 3.0 recalls per 1000 women screened with FFDM). The authors proposed that DBT recall rates may decline in Australia as screen readers become more experienced with the technique and previous rounds of DBT images are made available (which is the current status of FFDM).

Systematic reviews

We identified two systematic reviews with pooled analysis of overall recall rate (Marinovich et al., 2018; Phi et al., 2018). A write-up of the pooled analysis methodology is provided in *section 3.1.1*. Results from the pooled analysis suggested that, in programs with double reading and biennial screening, recall rates did not decrease with the use of DBT (but recall rates did decrease for annual screening with single reading protocols).

Results from the Marinovich et al.'s meta-analysis reported a mixed result for overall recall rate when DBT was used as part of the screening test. Marinovich et al.'s 2018 pooled analysis of data from European trials reported an overall pooled estimate showing a statistically significant absolute decrease in recall rate for DBT compared with FFDM (–2.2 percent; 95%CI: –3.0, –1.4, p<.001, I2=98.2 percent). Stratified analyses however indicated a difference between paired and unpaired studies, noting that decrease to be mainly attributable to unpaired studies (pooled difference in recall rate = –2.9 percent; 95%CI: –3.5, –2.4, p<.001, I2=92.9 percent), with no statistically significant difference in recall

Study Design	Results: overall recall rate (95%CI)
Marinovich et al., 2018 Meta-analysis including data from Malmö, OTS and STORM	Recall rate for paired studies: FFDM+DBT: 0.5% increase (-0.1, 1.2) compared to FFDM alone (p=.12) Recall rate for unpaired studies: FFDM+DBT: 2.8% decrease (-3.5, -2.1) compared to FFFDM alone (p<.001) All -1.8% decrease (-2.7, -0.9, p<.001)
Phi et al., 2018 Systematic review with meta-analysis including data from Malmö and STORM	Recall rate for three paired studies: Pooled RR=1.12 (0.76, 1.63) Recall rate for seven unpaired studies: RR=0.72 (0.64, 0.80)

rates observed for paired studies (0.5 percent increase in recall; 95%CI: –0.1, 1.2, *p*<.12; I2=93.5 percent; *p*<.001 for difference between strata). The paired studies are more similar to the BSA program. Given the significant heterogeneity in results in the primary studies and between the studies operating in contexts more (or less) like that seen in Australia, caution is recommended in relation to the results. These differences are likely to reflect underpinning differences in the screening process (eg, single reading and an annual screening interval is used in most of the retrospective analysis, but the paired trials used biennial screening and double reading).

Phi et al. (2018) completed a systematic review with meta-analysis to understand the relationship between DBT compared to FFDM for women with dense breasts. All papers were included in *Allen + Clarke*'s 2018 review on DBT in screening and are subject to the same differences in study design, intervention/ comparator and screening interval as described for Marinovich et al.'s study. The pooled estimates from three paired European studies indicated that FFDM+DBT did not reduce recall rate (I2=76 percent). The pooled RR was 1.12 (95%CI: 0.76, 1.63). Subgroup analysis was not performed due to a small number of studies. Pooled estimates for seven US studies showed a significant reduction in recall rate when using DBT+/- FFDM compared to FFDM alone.

DBT+s2DM compared to FFDM+DBT and/or FFDM alone

Mixed results were also presented for the studies investigating DBT+s2DM compared to FFDM. RCT data suggested that recall decreases with DBT but other prospective studies set in screening environments suggest that recall did not decline.

RCT

In the To-Be-1 RCT (and its sub-studies), use of DBT+s2DM resulted in fewer recalls, which is consistent with the findings of the systematic reviews. In the main analysis, Hofvind et al. (2019) reported lower recall rates when women were screened with DBT+s2DM compared with FFDM. Recall rates for DBT+s2DM were 3.1 percent (95%CI: 2·8, 3·4) compared to 4·0 percent (95%CI: 3·7, 4·3; *p*<.0001). Hofvind et al. also reported that the recall rate increased during the study period for DBT (*p*=·01) and for FFDM (*p*=·01). Reasons proposed included that this finding reflected some natural variation, a study effect or, in the case of DBT, a learning effect.

A further sub-study (interim analysis) from this RCT was completed by Aase et al. (2019). They assessed cases discussed at consensus (a meeting between two or more radiologists to discuss all findings identified as probably benign or suspicious of cancer) and recall rates by breast density (using a Volpara density grade, VDG) and by prevalent or incidence screening round. In this study, fewer cases were discussed at consensus for DBT+s2DM compared to FFDM (6.4 percent compared to 7.4 percent, *p*=.03). Following consensus, fewer women were recalled when DBT+s2DM was used (3 percent with DBT+s2DM compared to 3.6 percent with FFDM). No differences were observed between women undergoing prevalent screening (i.e., there was no increased recall for women undergoing mammography for the first time). A difference was observed for incident screening, with fewer women recalled with DBT+s2DM. Decreases in recall were also observed for women screened with DBT who have less dense breasts (see *Table 15a*).

Prospective studies embedded in screening programs

Mixed results were presented in the prospective studies reporting on recall for DBT+s2DM compared to FFDM alone. Some studies reported decreasing recall with DBT, others (including

data from the Maroondah trial) reported increases in recall rate. Some, like the Verona pilot evaluation, observed no significant differences (Caumo, et al., 2018).

The Maroondah pilot reported by Houssami et al. (2019) reported recall rates for DBT+s2DM to be higher than for FFDM: DBT+s2DM had a recall rate of 4.2 percent (95%CI: 3.6, 4.8) compared with 3.0 percent (95%CI: 2.6, 3.5) for FFDM. The recall rates for women aged under 60 years and for those aged over 60 years were similar. Houssami et al. suggested that DBT recall rates may decline in Australia as screen readers become more experienced with the technique and have DBT screens from earlier screen rounds available for review and comparison.

Bernardi et al. (2019) reported data from the Trento pilot evaluation using an independent double reading strategy with consensus. Unlike the Maroondah trial, it demonstrated lower recall rates for DBT+s2DM. The study design used screening round data stratification (prevalent and incident) and observed that the reduction in recall observed by DBT+s2DM was mostly achieved through a reduction in recall for prevalent screening examinations. The study indicated that recall percent was generally lower for incident screening rounds for both DBT+s2DM and FFDM (around 2.3 percent). This may have been due to the availability of prior mammograms (as lower recall in incident screening is expected). The authors considered that, there was limited scope for DBT to further reduce the already low recall percentage.

A BreastScreen Norway study reported by Hofvind et al. (2018) which used an independent double reading strategy with consensus, found no significant difference in recall rate between DBT+s2DM (3.4 percent) and FFDM (3.3 percent, p=.563). Additionally, no significant differences were found for either prevalent screening (DBT+s2DM: 7.4 percent; FFDM, 7.6 percent; p=.591) or incident screening (DBT+s2DM: 2.4 percent; FFDM: 2.5 percent; p=.284).

Recall rates for the Cordoba trial (Romero Martín et al., 2018) were observed to be lower at 2.9 percent for single reading of DBT+s2DM (arm 3) compared with 3.1 percent for double reading of FFDM (arm 2). In the study a reduction in recalls of 40.5 percent (95%CI: 37.2, 43.9, p<.001) for single reading DBT+s2DM compared to double FFDM was reported. This study also reported a significant increase in cancer detection and an increased rate of PPV recalls and PPV biopsies for single reading DBT+s2DM. For this reason, the authors propose that single reading of DBT+s2DM could be an alternative to double reading of FFDM.

Retrospective analysis

Hovda et al. (2019) reported a decrease in recall rates with DBT compared to FFDM. Hovda et al. investigated the impact of subsequent screening with four screening protocols: FFDM after FFDM, DBT after FFDM, DBT after DBT, and FFDM after DBT. The study found that recall for women screened with two FFDM screening examinations was 3.6 percent, which was higher than for all other study groups (p<.001). The lowest recall rate (1.9 percent) was observed among women with two consecutive DBT examinations. The recall for FFDM after DBT (2.2 percent) was significantly lower than for DBT after FFDM (2.7 percent, p<.001). Hovda et al. proposed that a significantly lower recall for those with a prior DBT compared to those with a prior FFDM may indicate that the availability of prior DBT examinations at screen reading and/or consensus lowered the probability of recall. This was also one of the findings of the To-Be-1 RCT (that the availability of all prior mammograms may influence CDR). Additionally, suspicious findings on screening images could be explained by the superimposition of overlapping tissue when compared with prior DBT images, resulting in a negative screening interpretation. A learning effect as the radiologists were becoming more experienced with DBT should also be considered.

FFDM+DBT compared to FFDM alone

RCT

The Reggio Emilia RCT (Pattacini et al. 2018) used an independent double reading strategy and arbitration, noted a recall rate of 3.5 percent for both DBT+FFDM and FFDM. This result becomes noteworthy when compared to the other primary outcomes reported in favour of DBT which were 90 percent more cancers detected (8.6 per 1000 compared to 4.5 per 1000); PPV almost doubled (24.1 percent compared to 13.0 percent), and false positive recall rates decreased (30 per 1000 to 27 per 1000). These statistics indicate that maintaining a steady recall rate while improving other outcomes is an overall positive effect.

Prospective studies embedded in screening programs

Both OTS trial studies reported in this literature review used an independent double reading and arbitration consensus protocol. Skaane et al. (2019) and Skaane et al. (2018) observed mixed results in relation to recall rates. In Skaane et al. (2018), reported results favoured a significantly lower recall rate for FFDM+DBT: 3.4 percent (33.7 per 1000 screening examinations) for FFDM+DBT compared to 4.2 percent (42.3 per 1000 screening examinations) for FFDM alone in previous rounds. This was a difference of -8.5 per 1000 observed (95%CI: -11.3, -5.7, p<.001). However, in the second study, Skaane et al. (2019) reported on the final data of the trial. The observed recall rate was 3.4 percent for FFDM+DBT compared with 2.6 percent for FFDM. The study authors acknowledged that in their previous paper that recall rates for two previous rounds of screening using FFDM had an average recall rate of 4.2 percent. In this later study, the post-arbitration recall rate for FFDM arms A + B was 2.6 percent. The authors stated that it is likely that the availability of DBT images at consensus meetings allowed cases that would have been recalled according to FFDM findings alone to be dismissed. To avoid distortions made after the consensus meeting, they focused on pre-consensus interpretations in their analysis. It is therefore possible to consider that this may have impacted the differences in recall rates reported in the two studies.

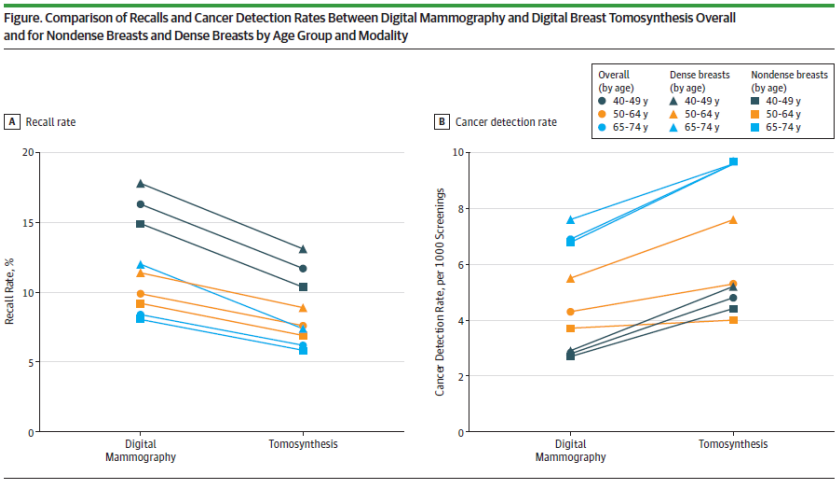
Observational studies: retrospective design

Four retrospective observational studies conducted in the US and one study conducted in the United Kingdom all reported lower recall rates for DBT+FFDM than FFDM alone.

The PROSPR consortium retrospective multicentre study conducted by Conant et al. (2019) found that recall rates for DBT compared to FFDM were lower. The generalized estimating equation analysis confirmed the lower recall rate for DBT (OR=0.64; 95%CI: 0.57, 0.72; p<.001) after adjustment for age group, breast density, first or subsequent screening, and research centre. As can be seen in *Figure 3* (overleaf), this pattern remained within every age group with similar OR across age groups. Recall rates were consistently higher for younger women, women with dense breasts, and those at first screening. Screening examinations among women with non-dense breasts had lower recall for DBT compared to FFDM (OR=0.62; 95%CI: 0.54, 0.72; p<.001), as did those among women with dense breasts (OR=0.65; 95%CI: 0.58, 0.73; p<.001).

Honig et al. (2019) undertook retrospective analysis of data from 1887 women (FFDM: 5029, DBT: 17,026) and identified that DBT+FFDM recorded a significantly lower recall rate than FFDM alone (8 percent; 10.6 percent; p<.001). Rose & Shisler also found that DBT compared to FFDM reduced

the recall rates. Rose and Shisler determined that for women younger than 50 years that recall rates with DBT+FFDM were 11.7 percent compared to 10.9 percent for FFDM alone (a difference of -7.4 percent, *p*=.003), and a non-significant reduction was observed for women with and without dense breasts (9.5 percent reduced to 8 percent [-15] and 13.6 percent to 13.2 percent [-4]) respectively.



One study also reported on reader recall results for women attending for prevalent screening in the United Kingdom's breast-screening program (Upadhyay et al., 2018). This was a very small study (compared to the other studies discussed in this literature review): only 880 women participated. A total of 153 women screened with FFDM were recalled and 101 women screened with FFDM+DBT were recalled (17.4 percent compared to 11.4 percent, an overall reduction of 35 percent). These recall rates seem very high overall, but the authors do not provide considerations as to why this might be. Recall was high regardless of breast density.

DBT alone verses FFDM alone

Prospective observational study

A Breast Cancer Surveillance study by Miglioretti et al. (2019) evaluated screening DBT performance by cumulative DBT volume. The timeframe was two years post-DBT adoption relative to FFDM performance one year before. The reader protocol was not reported. Miglioretti et al. reported that, before DBT adoption, FFDM recall rate was 10.4 percent (95%CI: 9.5, 11.4). After DBT adoption, it was lower (9.4 percent; 95%CI: 8.2, 10.6; *p*=.02). Relative to FFDM, DBT recall rate decreased for a cumulative DBT volume of fewer than 400 studies (OR=0.83; 95%CI: 0.78, 0.89) and remained lower as volume increased (OR=0.8; 95%CI: 0.75, 0.85; OR=0.81; 95%CI: 0.76, 0.87; OR=0.78; 95%CI: 0.73, 0.84]; OR=0.81; 95%CI: 0.75, 0.88; *p*=.001). Analysis was undertaken for radiologists who specialised in breast screening and for those who did not. Improvements were sustained for breast imaging subspecialists (OR range, 0.67, 0.85; *p*=.02) and readers who were not breast imaging specialists (OR range, 0.80, 0.85; *p*=.001). Recall rates decreased more in women with non-dense breasts (OR range, 0.68, 0.76; *p*=.001) than in those with dense breasts (OR range, 0.86, 0.90; *p*=.05). Miglioretti et al. observed an improvement in recall rate with no decrease in CDR, regardless of DBT volume, for both breast imaging subspecialists and readers who were not breast imaging subspecialists. Improvements in recall rates were also observed regardless of breast density but were larger in women with non-dense breasts than in those with dense breasts.

DBT_{MLO} compared to FFDM alone

Prospective studies embedded in screening programs

Results from the Malmö trial (Zackrisson et al., 2018) reported that lower recall rates with FFDM compared to DBT_{MLO} (2.5 percent compared to 3.5 percent, $p<.0001$).

3.4.2. False positive recall rate

False positive recall 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 her 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 have a benign final outcome. A reduction in false positive recall rate would be favourable for DBT as it would mean radiologists are making more accurate decisions about whether an abnormality seen on imaging requires investigation.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

Like overall recall rates, there was variance in the direction of effect when DBT was used, with some studies reporting a decrease in false positive recall rate and others reporting an increase (often associated with screening interval but not always). For example, Hodgson et al. (2016) reported that STORM and OTS trials observed different results for false positive rate when using FFDM+DBT compared to FFDM alone.

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

Results from STORM-2 also 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 (and noting that there is a learning effect associated with reading the synthesized images). 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.

Table 15a: DBT+s2DM compared to FFDM+DBT or FFDM alone: studies reporting on overall recall rate

Study	Sample	Study type	DBT+s2DM Overall recall rate (95%CI; p-value)	FFDM alone Overall recall rate (95%CI; p-value)	Difference between Overall recall rate detection (95%CI)
Randomized controlled trials					
Aase et al. (2019) To-Be-1 (sub-study)	14,274 women randomized to DBT (n=7155) or FFDM (n=7119) as part BreastScreen Norway biennial screening program.	First year of RCT imaged on GE SenoClair with independent double reading and consensus or arbitration. Non-dense breasts = VDG 1 and 2 Dense breasts = VDG 3 and 4	All: 3.0% Prevalent: 6.3% Incident: 2.3% VDG 1: 2.2% (1.4, 2.9) VDG 2: 2.5% (1.9, 3.0) VDG 3: 4.2% (3.3, 5.1) VDG 4: 3.6% (2.1, 5.2)	All: 3.6% Prevalent: 6.2% Incident: 3.1% VDG 1: 3.4% (2.5, 4.3) VDG 2: 3.6% (2.9, 4.2) VDG 3:3.8% (3.0, 4.7) VDG 4: 3.6% (2.1, 5.0)	All: p=.03 Prevalent: p=.95 Incident: p<.01 VDG 1: p=.04 VDG 2: p=.01 VDG 3: p=.56 VDG 4: p=.93
Hofvind et al. (2019) To-Be-1 (main analysis)	28,749 women aged 50-69y participating in BreastScreen Norway biennial screening program. DBT+s2DM: 14,380 FFDM alone: 14,369	Single site RCT of one screening round, imaged on GE SenoClair with independent double reading with independent double reading and consensus or arbitration.	3·1% (2·8,3·4)	4·0% (3·7, 4·3, p<.0001)	NA
Prospective studies embedded in population-based screening programs					
Bernardi et al. (2019) Trento	DBT+s2DM: 46,343 participants in the Trento biennial screening program (mean age = 57.9y) FFDM alone: historical cohort of 37,436 women previously screened in the Trento program (mean age = 57.9y)	Prospective single site pilot evaluation following the implementation of DBT into the Trento screening program imaged with Hologic systems. Independent double-reading with arbitration.	2.55%	3.21%	RR=0.79 (0.73, 0.86)

Study	Sample	Study type	**DBT+s2DM** Overall recall rate (95%CI; p-value)	**FFDM alone** Overall recall rate (95%CI; p-value)	**Difference between** Overall recall rate detection (95%CI)
Houssami et al. (2019) Maroondah	10,146 women presenting for biennial screening: DBT+s2DM: 4993 women (5018 screening examinations) FFDM alone: 5153 women (5166 screening examinations)	Prospective single site pilot trial embedded in BSA, imaged on Hologic Selenia Dimensions 8000 (+/- C-view software) or Siemens Mammomat Inspiration. Double reading with consensus or arbitration	All: 4.2% (3.6, 4.8) Prevalent: 7.6% (6.0, 9.4) Incident: 3.4% (2.8, 4.0) <60y: 4.2% (3.6, 5.0) >60y: 4.1% (3.3, 5.0)	All: 3.0% (2.6, 3.5) Prevalent: 7.4% (4.9, 10.7) Incident: 2.7% (2.2, 3.2) <60 y: 3.2% (2.4, 4.0) >60y: 2.9% (2.3, 3.5)	Estimated difference: All: 1.2% (0.46, 1.9) Prevalent: 0.12% (-3.4, 3.1) Incident: 0.69% (–0.02, 1.4) <60y: 1.1% (–0.01, 2.1) >60y: 1.2% (0.18, 2.3)
Caumo et al. (2018) Verona	DBT+s2DM: 16,666 participants in the Verona screening program FFDM: historical cohort of 14,423 women previously screened in the Verona program	Prospective single site pilot evaluation imaged with Hologic systems. Double reading with consensus or arbitration	All: 4.2% Prevalent: 6.4% Incident: 3.5%	All: 4.0% Prevalent: 7.6% Incident: 3.7%	All: *p*=.34 Prevalent: *p*=.15 Incident: *p*=.32
Hofvind et al. (2018) Norway BSP	DBT+s2DM: 37,185 attendees at an Oslo clinic for biennial (7250 had a prevalent screen), mean age=59.2y FFDM: 61,742 attendees at the Vestfold or Vestre Viken clinics (9517 had a prevalent screen), mean age=59.4y	Prospective, multi-site, population-based cohort study imaged on Hologic Dimensions, Siemens Mammomat Inspirations or GE SenoEssential units with double reading and consensus or arbitration.	All: 3.4% Prevalent: 7.4% Incident: 2.4%	All: 3.3% Prevalent: 7.6% Incident: 2.6%	All: *p*=.562 Prevalent: *p*=.591 Incident: *p*=.284
Romero Martín et al. (2018) Cordoba	16,067 women in the Cordoba biennial screening program (aged 57.59y) undergoing FFDM+DBT with s2DM images (3341 women attending for prevalent screening; 12727 for incident screening)	Prospective single site transversal reading study, imaged on Hologic Dimensions unit. Double reading without consensus or arbitration.	Third reading: 2.9% Fourth reading:2.8%* (FFDM+DBT+s2DM)	First reading= 3.5% Second reading = 3.1%	NA
Retrospective analysis					
Hovda et al. (2019) Norway BSP	35,736 women with at least two consecutive screens in the BreastScreen Norway biennial screening program	Retrospective paired analysis of data using Hologic Dimensions and GE Senographe systems. Double reading protocol with arbitration or consensus.	DBT after DBT = 1.9% DBT after FFDM= 2.7%	FFDM after FFDM=3.6% FFDM after DBT= 2.2%	FFDM after FFDM=3.6% FFDM after DBT= 2.2%

Table 15b: FFDM+DBT compared to FFDM alone: studies reporting on overall recall rate

Study	Sample	Study type	FFDM+DBT Overall recall rate (95%CI; p-value)	FFDM alone Overall recall rate (95%CI; p-value)	Difference between Overall recall rate **detection** (95%CI)
Randomized controlled trials					
Pattacini et al. (2018) Reggio Emilia	Women aged 45-74y years attending for incident screening (women aged 45-49y screened annually; women aged ≥50 screened biennially). Women with family history were excluded. FFDM: 9783 women (mean age=56.3y) FFDM+DBT: 9777 women (mean age=56.2y)	Baseline data from a prospective single site RCT using a test and treat methodology using GE Senographe Essential systems. Double reading with arbitration or consensus.	3.9%	3.7%	NA
Prospective studies embedded in population-based screening programs					
Skaane et al. (2019) OTS	24,301 women (59.1 years ± 5.7) Arm A = FFDM Arm B = FFDM+CAD Arm C = FFDM+DBT Arm D = s2DM+DBT	Prospective, single site, double-reading trial using Hologic Dimensions system, single compression per view. CAD system ImageChecker 9.3 (Hologic)	DM + DBT (arms C + D). 3.4% (820/24 301)	FFDM (arms A + B) 2.6% (623/24 301)	NA
Skaane et al. (2018) OTS	24,301 women aged presenting for screening in Oslo, Norway. Women aged 40-54y were screened every 18 months and women aged 55-74y biennially.	Prospective, single site fully paired, double-reading trial using Hologic Dimensions system (FFDM, DBT and s2DM images) and GE Senographe. Double reading strategy.	3.7% (820/24,301)	4.23% (2530/59,877)	− 8.5 (− 11.3, − 5.7; *p*<.001)
Retrospective analysis					
Bahl et al (2019)	35,665 women aged 65+ years and older. FFDM: 15,019 women (mean 72.7 years) DBT: 20,646 women (mean 72.1 years)	Retrospective analysis, single academic medical centre using Selenia Dimensions; Hologic. Single reader protocol.	AIR: 5.7% (5.4, 5.9)	AIR: 5.8% (5.6, 6.1)	*p*<.001

Conant et al. (2019) PROSPR consortium	96,269 women aged 40-74y participating in annual screening through the PROSPR consortium, mean age=55.7y FFDM: 129,369 screening examinations (mean age 56.4y) FFDM+DBT: 50,971 screening examinations (mean age 54.6y)	Retrospective, three site analysis with DBT images collected on Hologic Selenia Dimensions unit, varying DM units (not recorded) with single reading (n=2 radiologists)	First 40-49 y 50-64 y 65-74 y	All 28.5% 20.1% 16.3%	N/dense 28.8% 17.8% 16.5%	Dense 28.6% 24.7% 16.0%	First 40-49 y 50-64 y 65-74 y	All 17.1% 14.8% 13.9%	N/dense 15.9% 15.8% 12.4%	Dense 18.5% 13.1% 17.5%
			Subs- 40-49 y 50-64 y 65-74 y	14.4% 9.3% 8.1%	12.8% 8.8% 7.8%	15.9% 10.6% 9.2%	Subs 40-49 y 50-64 y 65-74 y	14.4% 7.1% 5.8%	9.0% 6.3% 5.5%	11.7% 8.6% 6.8%
Honig et al. (2019)	1887 women (22,055 screening mammograms) FFDM Vs DBT. FFDM: 5029 DBT: 17,026	Retrospective analysis. Three outpatient sites of one academic institution. Reader protocol not stated.	8%				10.6% (*p*=.001)			
Miglioretti et al (2019)	271,362 women (mean age=57.5y), from 2010 to 2017, interpreted by 104 radiologists from 53 facilities in the Breast Cancer Surveillance Consortium. DBT: 106 126 DBT examinations FFDM: 221 248 examinations	Multi-site prospective study. Reader protocol and units not stated.	DBT alone: All radiologists = 9.4% (8.2, 10.6) Sub-specialist yes = 9.7% (8.5, 11.1) no = 9.6% (8.1, 11.4)				FFDM alone: 1 year prior to DBT adoption All radiologists = 10.4 (9.5, 11.4) Sub-specialist yes = 10.0% (9.0, 11.1) no = 11.0% (9.8, 12.4)			
Rose & Shisler (2018)	59,921 screening examinations of women aged under 50y (no lower age range provided, no stratification), estimate of 10 percent attending for prevalent screening examination: FFDM: 41,542 examinations FFDM+DBT: 18,379 examinations	Multi-site (N=31) retrospective study set in community mammography practices imaged on Hologic Selenia Dimensions units. Reader protocol not stated.	Women <50y All: 10.9% Non-dense: 8.0% Dense: 13.2%				Women <50y All: 11.7% (*p*=.017) Non-dense: 9.5% (*p*=.6) Dense: 13.6% (*p*=.027)			

GRADE assessment: overall recall rate: FFDM+DBT compared to FFDM alone

Participants Studies	Quality of evidence	Overall results
217,565 participants 6 studies	⊕⊕ Low	Pooled analysis and data from prospective fully paired trials and retrospective analysis shows inconsistent results.

GRADE assessment: overall recall rate: DBT+s2DM compared to FFDM+DBT or FFDM alone

Participants Studies	Quality of evidence	Overall results
100,752 participants 3 studies	⊕⊕ Low	Pooled analysis and data from prospective fully paired trials and retrospective analysis shows inconsistent results.

GRADE assessment: overall recall rate: DBT_{MLO} compared to other imaging

Participants Studies	Quality of evidence	Overall results
7,681 2 studies	⊕ Very low	No systematic review or pooled analysis was available.

Updated findings

Literature published since 31 December 2017 relating to false positive recalls and DBT reported results for the Reggio Emilia RCT, STORM-2, OTS and Malmö trials along with three retrospective analyses. No systematic reviews of studies published since December 2017 considered false positive recall rate. The small number of studies reporting false positive results provided a wide cross-section of screening protocols including FFDM+DBT, DBT+s2DM and DBT_{MLO}. Different reading and consensus/arbitration protocols were reported, which might influence results quite considerably and one European study (perhaps surprisingly) did not use consensus or arbitration for discordant reader results. The methods sections of some studies lacked relevant detail including in a number of cases a definition of false positive recalls. Whether false positive recall rates for DBT are associated with learning curves, availability of prior screens or other methodological factors it is currently difficult to determine. Studies in progress may provide greater clarity on this important area in the future. *Table 17* includes a summary of all study results. Detailed methodologies for each of these studies is provided in *section 3.1.1*). Included studies are listed below.

Randomized controlled trial

One randomized study embedded in population-based screening: Pattacini et al., 2018

Prospective trials embedded in population screening programs

Three studies: Skaane et al., 2019; Bernardi et al., 2018; Zackrisson et al., 2018,

Observational studies

Three studies with a retrospective design: Bahl et al., 2019; Honig et al., 2019; Hovda et al., 2019

Summary of key findings published between 1 January 2018 and 31 December 2019

Inconsistent results regarding the influence of DBT on false positive recall rates were reported. Of the six studies that reported results, two prospective studies reported higher false positive recall rates for DBT+s2DM compared to FFDM and four studies reported lower rates of false positive recall. Studies that recorded increases in false positive recall rates for DBT were the STORM-2 trial and Malmö trial. The STORM-2 trial did not employ a means of arbitration for discordant screen readings which may have influenced the result (i.e., a lower false positive recall rate could have been achieved if third party consensus was used). Additionally, other factors that could have impacted higher false positive recall rates in this study suggested by the authors were learning effect in relation to adjusting to new imaging protocols that produce radically different images and the process of sequential screen-reading at the DBT phase which allowed within participant comparison. Other studies also noted a learning effect in relation to reader lack of familiarity with $DBT_{MLO.}$ Another factors to consider is the lack of availability of prior DBT screens that could be compared unlike FFDM. The inclusion of DBT or DBT+2SDM in population breast screening poses a difficult question in relation to the potential trade-off between increasing CDR but also increasing false positives.

DBT+s2DM compared to FFDM+DBT and FFDM alone

Mixed results are presented on the impact that DBT+s2DM has on false positive recall rate.

Prospective trials embedded in population screening programs

In a secondary analysis from the STORM-2 paired prospective trial, Bernardi et al. (2018) compared double reading for four reading strategy in two arms. Arm 1 involved sequential reading of FFDM alone then FFDM+DBT; Arm 2 involved independent sequential reading of the same screening examinations using s2DM alone then DBT+s2DM by two different radiologists. Given that four radiologists read each set of images, consensus or arbitration was not used. If one or both readers in a pair noted a positive screen, the woman was recalled for further investigation. This study determined a false positive rate for several modalities. A definition of false positive was not provided. Use of DBT resulted in an increase in the false positive recall rate:

- FFDM: range was 1.2 percent and 2.7 percent (median 2.25 percent)
- FFDM+DBT: range was 1.5 percent and 3.4 percent (median 2.75 percent), and
- s2DM: range was 1.6 percent and 4.6 percent (median 2.4 percent)
- DBT+s2DM: range was 1.8 percent and 6.7 percent (median 3.0 percent).

The study found that the integration of DBT in screen-reading increased the number and rate of false positive recall rate for most radiologists involved compared to FFDM or s2DM (see *Tables 16a* and *16b*). Bernardi et al. suggested that the sequential screen-reading used in the trial to

enable within-participant comparisons may have contributed to the higher false positive recall. Bernardi et al also highlighted that STORM represented the radiologists' first screening experience using s2DM images. They expected that increasing experience with s2DM will reduce false positive recall in future screening rounds; however, no further evidence on false positive recall within pilot evaluations at STORM-2 sites was reported by either Bernardi et al. (2019) or Caumo et al. (2018) (i.e., at the Trento or Verona pilot sites).

Table 16a: STORM-2 radiologist-specific false positive recall using FFDM and DBT

FFDM alone Readers	FFDM+DBT Readers	FP:TP [19]detection attributed to integrating 3D mammography in Screening
1= 13 (2.2%, 1.2, 3.8) 2= 83 (2.7%, 2.1, 3.3) 3= 102 (2.6%, 2.2, 3.2) 4= 52 (1.2%, 0.9, 1.5) 5= 106 (2.0%, 1.6, 2.4) 6= 40 (2.3%, 1.6, 3.1)	1= 20 (3.4%, 2.1, 5.3) 2= 86 (2.8%, 2.2, 3.4) 3= 118 (3.1%, 2.5, 3.6) 4= 67 (1.5%, 1.2, 1.9) 5= 145 (2.7%, 2.3, 3.2) 6= 39 (2.2%, 1.6, 3.0)	1= ↑7 : 0 2= ↑3 : ↑5 3= ↑16 : ↑8 4= ↑15 : ↑6 5= ↑39 : ↑19 6= ↓1 : ↑2

Table 16b: STORM-2 radiologist-specific cancer detection measures at screen-reading using s2DM and DBT

S2DM Readers	DBT+s2DM Readers	S2DM+DBT FP:TP [20]detection attributed to integrated 3D mammography
1= 29 (4.6%, 3.1, 6.6) 2= 66 (1.6%, 1.2, 2.0) 3= 81 (2.6%, 2.0, 3.2) 4= 50 (1.6%, 1.2, 2.2) 5= 78 (2.4%, 1.9, 3.0) 6= 67 (3.9%, 3.0, 4.9) 7= 63 (1.9%, 1.5, 2.5)	1= 42 (6.7%, 4.9, 9.0) 2= 76 (1.8%, 1.5, 2.3) 3= 105 (3.3%, 2.7, 4.0) 4= 58 (1.9%, 1.5, 2.5) 5= 97 (3.0%, 2.5, 3.7) 6= 69 (4.0%, 3.1, 5.0) 7= 71 (2.2%, 1.7, 2.8)	S2DM+DBT 1= ↑13 : ↑2 2= ↑10 : ↑7 3= ↑24 : ↑8 4= ↑8 : ↑6 5= ↑19 : ↑4 6= ↑2 : ↑1 7= ↑8 : ↑1

Retrospective analysis

In a retrospective analysis of a BreastScreen Norway study, Hovda et al. (2019) did not specifically report on false positive recall rates or provide a full definition of false positive ('women screened and recalled, but not diagnosed with breast cancer'). They noted in their analysis that as the 'PPV_1 was significantly higher among those screened with DBT; thus, the rate of false positives was lower'. This study had a complex reading protocol, and it is not clear whether their reference to DBT in this instance related to DBT+FFDM or DBT+s2DM. They did however utilise an independent double screening with consensus as part of their reader protocol which may have been relevant to their findings.

[19] Indicates the number of FP recalls increased (↑) or decreased (↓) to the additional number of breast cancers detected by integrating 3D with 2D mammography for breast screening.
[20] Indicates the number of FP recalls increased (↑) or decreased (↓) to the additional number of breast cancers detected by integrating 3D with synthesised 2D mammography for breast screening.

FFDM + DBT compared to FFDM alone

RCT

Baseline data from the Reggio Emilia RCT compared FFDM+DBT with FFDM (independent double reading with consensus). No definition of false positive was provided in their study write-up. This study reported a decrease in false positive recall rates from 3.0 percent with FFDM to 2.7 percent with FFDM+DBT (27 per 1000). No further analysis for false positive was completed; however, we note that the false positive recall rates between the two arms are very similar and also reflect the overall similarity between the two arms for recall rate (see *section 3.4.1*).

Prospective trials embedded in population screening programs

Skaane et al. (2019) reported false positive recall findings from the final data of the OTS trial. The study design reported on each screening examination as independently interpreted with the study's four reading modes: (A) FFDM, (B) FFDM + computer-aided detection (CAD), (C) FFDM+DBT, and (D) DBT+s2DM. For each reading mode, two false positive estimates were developed: pooled estimates based on all observations and reader-adjusted estimates (the average of radiologist-specific estimates). Statistical analysis was based on pre-arbitration reader-adjusted estimates (given the variance in the number of cases read by individual radiologists, which ranged from 564 cases to 5318 cases). Significantly lower false positives were found for FFDM+DBT compared to FFDM alone (p<.001).

- FFDM compared to FFDM+DBT: 5.8 (6.6) percent compared to 5.0 (5.4) percent[21]
- FFDM compared to FFDM+CAD: 5.8 (6.6) percent compared to 5.8 (6.8) percent[22]
- FFDM+DBT compared to DBT+s2DM: 5.0 (5.4) percent compared to 4.6 (5.1) percent[23]
- Simulated double reading of FFDM compared to FFDM+DBT: 9.7 (10.4) percent compared to 8.1 (8.4) percent[24].

Retrospective analysis

Bahl et al. (2019) reported on a single site retrospective study conducted in the United States which used single reading of consecutive screening mammograms in older women (65 years and older) comparing FFDM with FFDM+DBT across a range of performance measures. The FFDM screens were captured 2008 - 2011 (prior to integration of DBT) for 15 019 women (mean age of 72.7 years). The FFDM+DBT screens were captured from 2013 - 2015 (after complete DBT integration), for 20,646 women (mean age 72.1 years). The screens were interpreted by 29 breast imaging radiologists. Mammograms were considered false positive if there was no known tissue diagnosis of cancer within one year of a positive screening result (i.e., BIRADS reporting category 0, 3, 4, or 5). Reader protocol was not reported in this study (i.e. independent double or single). In relation to false positive recall rate, Bahl et al. reported that FFDM+DBT produced a lower false

[21] FFDM compared to FFDM+DBT: Reader-adjusted difference -1.2% (95%CI: -1.7,-0.7%; *p*<.001).
[22] FFDM compared to FFDM+CAD: Small and not significant (not stated).
[23] FFDM+DBT compared to DBT+s2DM: Reader-adjusted difference 1.0%(95% CI: -6.2%, 8.5%; *p*=.77)
[24] FFDM Double compared to FFDM+DBT: Reader-adjusted difference 4.7%, 5.0% and 5.0% respectfully and a difference of -0.3% (95%: -0.8%, 0.2%; *p*=.23)

positive rate compared to FFDM (4.8 percent compared to 5.1 percent; AOR, 0.85; *p*=.001). They also found that the potential risks of false positive examinations were lower with increasing age (AOR 70-74 yrs.; 0.94, 75-79 yrs.; 0.92, 80-84 yrs.;0.83, and 85 and over; 0.76, trending *p*<.01).

Honig et al. (2019) conducted a retrospective study reporting on 22,055 screening mammograms (1278 mammograms where the woman was recalled), to better understand some of the factors that influence recall rate with DBT and FFDM (including false positive recalls) (methodology described in *section 3.4.1*). Honig et al. defined false positive as BIRADS reporting category 1 or 2 at diagnostic imaging with one-year cancer-free follow-up. Details of the reader protocol were not provided. The study did not compare FFDM to DBT directly, but we have included this study here because it provides some useful insights into factors that might influence the false positive recall rate. Information about mammography (rather than other clinical factors) reported by Honig et al. included that false positive results were significantly lower if prior mammograms were available (90.8 percent compared to 95.8 percent; *p*=.02) but it did not matter whether this mammogram was FFDM or DBT. The study also found that there were no significant differences in false positive recall rates based on history of high-risk lesions, family history of breast or ovarian cancer, hormone use, breast density, race, or body mass index. Age however was an indicator for difference in that false positive recall rates were highest among women aged 40-49 years with rates decreasing with increasing age. Honig et al. proposed that these findings are relevant to practice in that they may influence decision for consideration of double reading (when it is not standard procedure) for women aged 40-49 years, who have less radiological images to compare against. This is of limited relevance to the BSA program however as all images are double read (and there was no strong evidence that a single reading of a DBT image is a clinically appropriate approach at this time).

DBT_{MLO} compared to FFDM

Prospective trials embedded in population screening programs

In the Malmö trial, Zackrisson et al. (2018) the authors determined that the proportion of false positive results after the consensus meeting was higher with DBT_{MLO} screen-reading than with FFDM. No data was presented but a graphic indicated that, at Year One, false positive recall rate was approximately 2.5 per 100 women screened for DBT_{MLO} compared to approximately 0.6 per 100 women screened with FFDM. Zackrisson et al. reported that the proportion of false positive results among recalled participants decreased in the DBT_{MLO} group during the first year of the trial and then stabilised. By Year Five, false positive recall rates were lower for DBT_{MLO} (approximately 1.5 per 100 women screened) but were still higher than FFDM alone (approximately 0.8 per 100 women). This may be as a result of learning effect when using $DBT_{MLO,}$ but no further commentary is provided about the continuing higher false positive rate or its significance within this trial (beyond noting a potential prevalence effect with the increased sensitivity of DBT as an imaging modality). Zackrisson et al. reported that they are planning to conduct further analysis to better understand the false positive results reported in the final Malmö dataset.

Table 17: Studies reporting on false positive recall rate

Study	Sample	Study type	FFDM + DBT False positive recall rate (95%CI; p-value)	DBT + s2DM False positive recall rate (95%CI; p-value)	FFDM alone False positive recall rate (95%CI; p-value)	Difference between false positive recall rates detection (95%CI)
RCT						
Pattacini et al. (2018) Reggio Emilia	Women aged 45-74y attending for incident screening	Baseline data from a prospective single site RCT using a test and treat methodology using GE Senographe Essential systems	All: 2.7% 45-49y: 3.4% 50-59y: 2.3% 60-70y: 2.6%	NA	All: 3.0% 45-49y: 3.4% 50-59y: 3.0% 60-70y: 2.7%	NA
Prospective studies embedded in population-based screening programs						
Bernardi et al. (2018) STORM	9672 asymptomatic Italian women aged 53-63y (median age 58y) who attended population-based screening	Prospective, fully paired trial using Selenia Dimensions system with C-view for FFDM + DBT with single reading	FFDM+ DBT: median 2.75% (1.5, 3.4)	DBT+s2DM: median 3% (1.8, 6.7) s2DM alone: median 2.4% (1.6, 4.6)	FFDM: median 2.25% (1.2, 2.7)	NA
Retrospective analysis						
Bahl et al. (2019)	35,665 women aged 65+y. FFDM: 15,019 (mean=72.7y) DBT: 20,646 (mean=72.1y)	Retrospective analysis, single academic medical centre using Selenia Dimensions; Hologic. 29 breast imaging radiologists	4.8% (4.6, 5.1)	NA	5.1% (4.9, 5.4)	AOR= 0.85 (0.78, 0.92, *p*<.001)
Honig et al. (2019)	1887 women (22,055 screening mammograms) FFDM Vs DBT. FFDM: 5029 DBT: 17,026	Single site retrospective analysis. False positive estimate	Estimate: 7.5%	Estimate: 10.0%		

3.4.3. Relative specificity

The specificity of a screening test refers to the proportion of women without breast cancer correctly identified by a test or the true negative rate. Increased specificity with DBT (compared to FFDM) indicates that use of DBT correctly identifies women who do not have breast cancer.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

Most study results demonstrated an improvement in specificity with the addition of DBT. Hodgson et al. (2016) reported an overall increase in specificity from the STORM trial for FFDM+DBT (96.49 percent specificity; 95%CI: 96.04, 96.90) compared with FFDM alone (95.55 percent specificity; 95%CI: 95.04, 96.01). Other reported results did not achieve statistical significance. 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 could be slightly higher for FFDM+DBT compared to FFDM, but again, the result was not significant (p=.777).

Updated findings

Six studies reported on relative specificity rates, providing a little more information about DBT's relative specificity compared to imaging with FFDM. Included studies are listed below: most studies reported on specificity for FFDM compared to FFDM+DBT (which is different from the balance of imaging modalities reported on in other parts of this literature review). Detailed methodologies for each of these studies is provided in *section 3.1.1*). *Table 18* includes a summary of all study results.

Prospective trials embedded in population screening programs

Two studies (three papers): Skaane et al., 2019; Skaane et al., 2018; Zackrisson et al., 2018

Observational studies

Three studies with a retrospective design: Bahl et al., 2019; Conant et al., 2019; Hovda et al., 2019

Summary of key findings published between 1 January 2018 and 31 December 2019

Studies reported an increase in relative specificity when DBT was used. Most studies reported relative specificity as greater than 91 percent for DBT imaging, and greater than 88 percent for FFDM imaging. Increases were generally very small, ranging from a 0.5 percent increase to 2.4 percent (which indicates FFDM's good performance for this metric, as well as DBT's superior performance). Only one of these studies (data from the OTS trial) was set in a screening program with similar parameters to the BSA program. The study which did not see improved specificity adopted DBT_{MLO} and suggested that an improved specificity would likely be seen with the increase of a second DBT view.

DBT+s2DM compared to FFDM+DBT and/or FFDM alone

Retrospective analysis

Hovda et al. (2019) completed a retrospective study comparing women who had two consecutive screens with FFDM to women whose first screen was FFDM and the subsequent screen was either FFDM+DBT or DBT+s2DM. This study determined that specificity was very high with both imaging modalities but that the group of women screened with FFDM first followed by a subsequent DBT screening examination (either FFDM+DBT or DBT+s2DM) was higher:

- Group 1 (two FFDM screens): 96.9 percent, compared to
- Group 2 (FFDM then FFDM+DBT or DBT+s2DM): 98.0 percent.

FFDM+DBT compared to FFDM alone

Prospective trials embedded in population screening programs

Two papers reported on program specificity in the OTS trial, a prospective study embedded in a population breast screening program. The first paper reported by Skaane et al. (2018) compared FFDM+DBT with the outcomes of two prior rounds of FFDM. Study findings indicated that specificity improved with the use of FFDM+DBT. The difference in specificity was a significant increase with FFDM+DBT reaching 97.5 percent specificity and FFDM alone reaching 96.4 percent (p<.001).

In the second paper, final results of the OTS trial were reported by Skaane et al. (2019). This paper reported on each screening examination as independently interpreted with the study's four reading modes: (A) FFDM, (B) FFDM + CAD, (C) FFDM+DBT, and (D) DBT+s2DM. For each reading mode, two specificity estimates were developed: pooled estimates based on all observations and reader-adjusted estimates (the average of radiologist-specific estimates). Statistical analysis was based on pre-arbitration reader-adjusted estimates (given the variance in the number of cases read by individual radiologists, which ranged from 564 cases to 5318 cases). Specificity was higher when DBT was used:

- FFDM compared to FFDM+DBT: 94.2 percent compared to 95.0 percent[25]
- FFDM compared to FFDM+CAD: 94.2 percent compared to 94.2 percent[26]
- FFDM+DBT compared to DBT+s2DM: 95.0 percent compared to 95.4 percent[27]
- Simulated double reading of FFDM compared to FFDM+DBT: 90.3percent compared to 91.9 percent[28].

While the specificity rates for the OTS trial differ slightly between the two studies reported here, the result remains the same: significantly higher rates of specificity can be found for FFDM+DBT

[25] Reader-adjusted difference in FPF of 21.2 percent (95%CI: 21.7, 20.7; p<.001).
[26] Reader-adjusted difference in FPF 0.2 percent (95%CI: 20.3-0.7; p=.47).
[27] Reader-adjusted difference in FPF of 20.3 percent (95%CI: 20.8, p=.23).
[28] Reader-adjusted difference of 21.9% (95%CI: 22.5, 21.3; p<.001).

compared with FFDM alone, irrespective of double or single reading. The use of DBT+s2DM showed no statistically significant differences in specificity compared with FFDM+DBT but the overall percentage was marginally higher than that reported for FFDM; however, the authors did not compare FFDM to DBT+s2DM.

Retrospective analysis

In an American observational study (Conant et al., 2019), the authors reported that screening examinations with DBT were associated with significantly higher specificity (OR=1.46; 95%CI: 1.30, 165; *p*<.001) after adjustment for research centre, age group, and breast density. Use of DBT was also associated with significantly higher specificity in every age group and at each level of breast density (all *p*<.001). A second retrospective study in the US conducted by Bahl et al. (2019) also saw improved specificity to a level of significance with FFDM+DBT compared to FFDM alone (95.8 percent compared to 95.1percent; AOR=1.23; *p*<.001).

DBT_{MLO} compared to FFDM

Prospective trials embedded in population screening programs

The Malmö trial reported that when comparing DBT_{MLO} with FFDM alone, DBT_{MLO} had lower specificity (97·2 percent; 95%CI: 97·0, 97·5) compared with FFDM (98·1 percent; 95%CI: 97·9, 98·3) (Zackrisson et al., 2018). Zackrisson et al. proposed that the addition of a second DBT view would likely improve the specificity slightly.

Table 18: Relative specificity

Study	DBT+s2DM (95%CI; p-value)	FFDM+DBT (95%CI; p-value)	FFDM alone (95%CI; p-value)	DBT_{MLO} (95%CI; p-value)
Prospective studies embedded in population-based screening programs				
Skaane et al. (2019) OTS	NA	Single read: 95.0% Double read: 91.9%	Single read: 94.2% Double read: 90.3%	NA
Skaane et al. (2018) OTS	NA	97.5	96.4 (DBT higher: 1.2% (0.91, 1.40; *p*<.001)	NA
Zackrisson et al. (2018) Malmö	NA	NA	98·1% (97·9, 98·3)	97·2% (97·0, 97·5)
Retrospective analysis				
Bahl et al. (2019)	NA	95.1% (94.9, 95.3)	94.8% (94.6, 95.1)	NA
Conant et al. (2019) PROSPR	NA	*All*: 91.3% 40-49y: non-dense 89.6%; dense 86.0% 50-64y: non-dense 93.2%; dense 94.4% 65-74 y: non-dense 94.4%; dense 93.9%	All 88.9% 40-49y: non-dense 84.9%; dense 82.3% 50-64y: non-dense 91.1%; dense 89.1% 65-74y: non-dense 92.2%; dense 90.6%	NA
Hovda et al. (2019) BSP Norway	Group 2 (FFDM then FFDM+DBT or DBT+s2DM): 98.0%		96.9%	NA

3.4.4. Positive predictive value

Positive predictive value (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.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

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 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 indicated 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.

GRADE assessment: overall recall rate: FFDM+DBT compared to FFDM alone

Participants Studies	Quality of evidence	Overall results
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.

GRADE assessment: overall recall rate: DBT +s2DM compared to FFDM+DBT or FFDM alone

Participants Studies	Quality of evidence	Overall results
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.

GRADE assessment: overall recall rate: DBT_{MLO} compared to other imaging

Participants Studies	Quality of evidence	Overall results
0	Not reported	No studies reported on PPV_{1-3}.

Updated findings

Literature published since 31 December 2017 reported on PPV_{1-3}; however, there were some inconsistencies observed between the definitions for PPV_2 and PPV_3, with both being reported as the percentage of breast cancer detected from needle biopsy after recall.

For the purposes of this paper the following definitions will be used:

- PPV_1: the percentage of breast cancer cases detected among recalled women
- PPV_2: the percentage of recalled women recommended to have needle biopsy
- PPV_3: the percentage of breast cancer detected from needle biopsy after recall.

If a study has reported PPV_2 as the percentage of breast cancer detected from biopsy after recall, we have reported this as PPV_3. No systematic reviews assessing false positive recall rates have been completed. Included studies are listed below. Detailed methodologies for each of these studies is provided in *section 3.1.1*). *Table 19* includes a summary of all study results.

Randomized controlled trials

One RCT (one paper): Hofvind et al., 2019

One randomized study embedded in a population screening program: Pattacini et al., 2018

Prospective trials embedded in population screening programs

Six studies (seven papers): Bernardi et al., 2019; Skaane et al., 2019; Bernardi et al., 2018; Caumo et al., 2018; Hofvind et al., 2018; Romero Martín et al., 2018; Skaane et al., 2018

Observational studies

Four studies with a retrospective design: Bahl et al., 2019; Conant et al., 2019; Hovda et al., 2019; Rose & Shisler, 2018

Summary of key findings published between 1 January 2018 and 31 December 2019

There is strong, consistent evidence that use of DBT increases PPV_1 (i.e., DBT is a more accurate test compared to FFM alone. Usually, studies reported that PPV_1 approximately doubled but RCT evidence suggested a slightly lower increase (about 30 percent). PPV_3 also increased significantly with the use of DBT.

DBT+s2DM compared to FFDM+DBT and/or FFDM alone

RCT

The To-Be RCT conducted by Hofvind at al. (2019) reported on PPV_1 noting significantly higher PPV_1 among those screened with DBT+s2DM compared to those screened with FFDM alone:

- DBT+s2DM: PPV_1 was 21.4 percent (95%CI: 17·6, 25·2)
- FFDM alone: PPV_1 was 15·2 percent (95%CI: 12·3, 18·2; p=.011).

PPV_3 for DBT+s2DM was 37·7 percent compared to 32·1 percent for FFDM, which was higher but not significantly higher (p=.18).

Prospective trials embedded in population screening programs

All studies reported an increase in PPV_1 with the use of DBT, with PPV_1 usually doubling.

The Trento and Verona evaluation pilots reported a near doubling of PPV_1 rates when DBT+s2DM was used compared to FFDM. The Trento prospective study reported a difference of PPV_1 for DBT+s2DM at 34.07 percent compared with 17.07 percent for FFDM (Bernardi et al., 2019). The Verona prospective study reported PPV_1 for DBT+s2DM at 23.3 percent compared with 12.9 percent for FFDM (RR=1.81; 95%CI: 1.34, 2.47) (Caumo et al., 2018). Increases in PPV were also observed between prevalent and incident screening rounds in the Verona pilot evaluation.

Hofvind et al. (2018) also reported increased PPV_1 when DBT+s2DM was used: 27.8 percent compared to 18.6 percent with FFDM alone (p=.001). This was also reflected in another BreastScreen Norway analysis of consecutive screening rounds (Hovda et al., 2019). The results of this study found PPV_1 was 12.9 percent with two consecutive FFDM screens but 43.5 percent for women with two consecutive DBT screens (p<.001). PPV_1 was also higher for FFDM after DBT compared with FFDM after FFDM (p<.05), suggesting that DBT significantly improves PPV_1.

The Cordoba study conducted by Romero Martín et al. (2018) undertook a slightly different screening protocol although the addition of DBT saw an increase in PPV_1 and PPV_3. Their analysis involved a comparison of paired double reading of FFDM (PPV_1 was 9.4 percent) and a single reading of DBT+s2DM (PPV_1 was 18 percent, an increase of 47.8 percent, p<.001).

In relation to studies that compared DBT+s2DM with FFDM for PPV_3, inconsistent results were reported (which aligns to the PPV_3 reported in the To-Be-1 RCT):

- Hofvind et al. (2018) reported a PPV_3 of 53.7 percent with DBT+s2DM compared to 38.6 percent with FFDM (p<.001)
- Romero Martín et al. (2018) reported a non-significant increase in PPV_3 of 14.4 percent DBT+s2DM compared to 0.8 percent for double reading with FFDM, p=.961).

Hofvind et al. (2018) also provided biopsy rates for both screening protocols, indicating that biopsy rates were slightly lower for DBT+s2DM (1·8 percent; 95%CI:1.5, 2.0) than for FFDM (1.9 percent; 95%CI: 1.7, 2.1) which was not significant (p=·40).

FFDM+DBT compared to FFDM alone

RCT

Baseline PPV was reported for the Reggio Emilia RCT (Pattacini et al., 2018). Like other studies, the authors reported an increase in PPV_1 and, again, this increase was almost double that reported for FFDM alone. PPV_1 for FFDM+DBT was 24.1 percent compared to 13.0 percent for FFDM alone (p=.0002). The increased PPV_1 was observed across all age cohorts (see *Table 19* for data).

Prospective trials embedded in population screening programs

Skaane et al. (2018) reported updated data from the OTS trial. PPV_1 was only reported for FFDM+DBT but it was quite high (27.7 percent). They also reported only on PPV_3 for FFDM+DBT (55.0 percent), which seems very high. A second paper in 2019 (Skaane et al., 2019) presented

final results of the OTS trial. Skaane et al. (2019) reported the predictive value of a positive score estimated as the fraction of true-positive cases out of all cases with positive scores, reader-adjusted estimates of sensitivity and specificity combined with the estimated cancer rate (1.16 percent). The study determined that the predictive value of positive scores improved with the addition of DBT from 9.9 percent to 14.1 percent. Similar gains were observed for simulated double-reading modes with the addition of DBT.

Observational studies: retrospective design

Three retrospective observational studies conducted in the US that compared FFDM+DBT with FFDM alone provided PPV_1 and PPV_3 results, with one study also providing a result for PPV_2.

PPV_1 values for the three studies included:

- Conant et al. (2019): PPV_1 was significantly greater for FFDM+DBT (6.29 percent) compared to FFDM (3.85 percent; OR 2.00; 95%CI: 1.47, 2.72; p<.001).
- Bahl et al. (2019): PPV_1 increased with FFDM+DBT with the addition of DBT to FFDM (15.0 percent compared to 11.5 percent; AOR=1.19; p=.02).
- Rose & Shisler (2018): in women aged younger than 50 years, PPV_1 increased from 1.6 percent to 2.3 percent with the use of FFDM+DBT (a significant increase of 0.7; 95%CI: 0.04, 1.4; p=.04). For women with dense breasts, PPV_1 increased from 1.6 percent to 2.5 percent with FFDM+DBT (a non-significant difference of 0.9; 95%CI: 0.04, 1.8; p=.55).

The results for PPV_2 provided by Bahl et al. (2018) reported that the percentage of recalled women recommended to have needle biopsy was greater for FFDM+DBT at 58.8 percent (95%CI: 54.7, 62.7) compared to FFDM alone (56.4 percent; 95%CI: 51.4, 61.4). They also reported a higher PPV_3 rate of 61.5 percent (95%CI: 57.4, 65.5, p=.55) for FFDM+DBT compared with 59.6 percent (95%CI: 54.4, 64.6; p=.69) for FFDM alone but neither of these differences were significant (unlike that of PPV_1 in the same study).

Conant et al. (2019) and Rose & Shisler (2018) also observed increases in PPV_3 for FFDM+DBT compared to FFDM alone but these results did not reach significance (see *Table 19* for data). Both Conant et al. and Rose & Shisler also recorded increases in biopsy rate for DBT+FFDM compared to FFDM alone. Conant et al. recorded a rate of 28.08 per 1000 screening examinations for DBT+FFDM compared with 18.96 per 1000 screening examinations for FFDM alone (OR=1.23; 95%CI: 1.08, 1.40; p=.002). The increased biopsy rate observed by Rose & Shisler was 16.6 percent for FFDM+DBT compared with 13.5 percent for FFDM alone (p=.003). This was comparable among women with non-dense breasts (OR=1.29; 95%CI: 1.05, 1.58; p=.01) and women with dense breasts (OR=1.17; 95%CI: 1.00, 1.37; p=.06).

DBT_{MLO} compared to other imaging combinations

Prospective trials embedded in population screening programs

Zackrisson et al. (2018) reported a PPV_1 of 24.1 percent (95%CI: 20.5, 28.0) for DBT_{MLO} compared with 25.9 percent (95%CI: 21.6, 30.7) for FFDM. The addition of a second DBT view may have

improved the PPV_1 for the DBT study group. This was the only study to report a lower PPV_1 with the use of DBT and the single view may account for the difference.

Table 19: PPV

Study	DBT+s2DM (95%CI)	FFDM+DBT (95%CI)	FFDM alone (95%CI)	DBT_{MLO} (95%CI)	Difference (95%CI)
DBT+s2DM compared to FFDM alone					
RCT					
Hofvind et al. (2019) To-Be-1 (main analysis)	PPV_1: 21.4% (17.6, 25.2) PPV_2: 37.7% (31.7, 43.7) PPV_3: 1.8% (1.5, 2.0)	NA	PPV_1: 15·2%, (12.3, 18.2, *p*=.011) PPV_2: 32·1% (26.5, 37.7, *p*=.18) PPV_3: 1.9% (1.7, 2.1, *p*=.40)	NA	NA
Prospective studies embedded in population-based screening programs					
Bernardi et al. (2019) Trento	PPV_1: 34.07%	NA	PPV_1: 17.07%	NA	PPV_1 % ratio: 2.00; (1.69±2.36)
Caumo et al. (2018) Verona	PPV_1 All: 23.3% Prevalent: 18.1% Incident: 25.5%	NA	PPV_1 All: 12.9% Prevalent: 10.7% Incident: 13.5%	NA	NA
Hofvind et al. (2018)	PPV_1: 27.8% PPV_3: 53.7%	NA	PPV_1: 18.6%, *p*=.001 PPV_3: 38.6%, *p*=.001	NA	NA
Romero Martín et al. (2018) Cordoba	PPV_1: 18.0% PPV_3: 46.0%	NA	PPV_1: 13.2% PPV_3: 44.3%	NA	NA
Bernardi et al. (2018) STORM-2	PPV_1: 34.07%	NA	PPV_1: 17.07%	NA	PPV% ratio 2.00 (95%CI: 1.69±2.36)
Retrospective analysis					
Hovda et al. (2019)	PPV_1 (DBT after FFDM): 36.4%	NA	PPV_1 (FFDM after FFDM): 12.9%, *p*<.001	NA	NA
FFDM+DBT compared to FFDM alone					
RCT					
Pattacini et al. (2018) Reggio Emilia	NA	*PPV_1* All: 24.1% 45-49 years: 13.4% 50-59 years: 28.5% 60-70 years: 29.1%	*PPV_1* All: 13.0% 45-49 years: 7.8% 50-59 years: 12.8% 60-70 years: 21.1	NA	NA
Prospective studies embedded in population-based screening programs					
Skaane et al. (2018)	NA	PPV_1: 27.7% PPV_3: 55.0%	NA	NA	NA

Study	DBT+s2DM (95%CI)	FFDM+DBT (95%CI)	FFDM alone (95%CI)	DBT_{MLO} (95%CI)	Difference (95%CI)
Retrospective analysis					
Bahl et al. (2019)	NA	PPV_1: 14.5% (13.1, 15.9) PPV_2: 58.8% (54.7, 62.7) PPV_3: 61.5% (57.4, 65.5)	PPV_1: 11.9% (10.4, 13.4) PPV_2: 56.4% (51.4, 61.4) PPV_3: 59.6% (54.4, 64.6)	NA	NA
Honig et al. (2019)	NA	PPV_1: 5.9% PPV_3:38.3%	PPV_1: 5.1% PPV_3: 36.5%	NA	NA
Rose & Shisler (2018)	NA	PPV_1 All: 2.3% Dense: 2.5% PPV_3 All: 16.1% Dense: 17.2%	PPV_1 All: 1.6% Dense: 1.6% PPV_3 All: 13.8% Dense: 13.4%	NA	*PPV_1* All: 0.7 (0.04, 1.4; *p*=.04) Dense: 0.9 (0.04, 1.8; *p*=.55) PPV_3 All: 1.5 (−3.3, 6.3; *p*=.55) Dense: 3.2 (−2.8, 9.2; *p*=.29)
FFDM+DBT compared to FFDM alone					
Prospective studies embedded in population-based screening programs					
Zackrisson et al. (2018) Malmö	NA	NA	PPV_1 : 25·9% (21·6, 30·7) NPV: 99·6% (99·4, 99·7)	PPV_1 : 24·1% (20·5, 28·0) NPV: 99·8% (99·7, 99·9)	NA

3.5. Radiation dose and safety

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 sufficient radiation to acquire clear images is a concern with all such procedures. 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, false positive recall rates and reduced mortality from breast cancer) outweigh the increased risk from lifetime exposure to radiation for women participating in screening.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

Radiation dose varies with the image acquisition process used (DBT+s2DM, FFDM alone or FFDM+DBT), 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 was based on dual acquisition protocols (i.e., FFDM+DBT

compared to FFDM alone or DBT+s2DM compared to FFDM+DBT). Using average breast thickness and compression, the radiation dose required to acquire acceptable images with FFDM+DBT is approximately double that of FFDM alone (2.98mGy compared to 1.49mGy). DBT+s2DM usually reported a lower mean glandular dose (MGD) compared to FFDM+DBT but it was still higher than FFDM alone. Having FFDM+DBT as the preferred screening strategy has 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 aged 40.

For FFDM+DBT, 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 alone compared with FFDM was reported to be lower: DBT_{MLO} has about 70 percent of the MGD compared to FFDM alone.

Use of s2DM significantly reduces the effective dose of combined FFDM+DBT because the 2D image is synthesised not acquired (meaning no additional radiation beyond that needed to acquire the DBT images), making it more comparable to FFDM alone but with the improved detection rates associated with DBT. Initial studies indicated that the quality of images reconstructed from s2DM is acceptable, but further evidence was required to ensure that it can be used to accurately interpret microcalcifications and that it is at least is not inferior to FFDM.

GRADE assessment: overall recall rate: FFDM+DBT compared to FFDM alone

Participants Studies	Quality of evidence	Overall results
18,926 participants 4 studies	⊕⊕ Low	No systematic review or pooled analysis was available.

GRADE assessment: overall recall rate: DBT+s2DM compared to FFDM+DBT or FFDM alone

Participants Studies	Quality of evidence	Overall results
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) indicated that DBT+s2DM radiation dose is considerably less than that used to acquire FFDM images.

GRADE assessment: overall recall rate: DBT_{MLO} compared to other imaging

Participants Studies	Quality of evidence	Overall results
0	Not reported	No studies reported on PPV_{1-3}.

Updated findings

We want to know, based on current evidence, what the MGD readings are when DBT is used in 'real world' screening scenarios. Literature published since 31 December 2017 reported on MGD using a combination of approaches and systems and in a wider range of 'real world' settings compared to what was available in late 2017. More data on a range of different imaging protocols

is available, especially for DBT+s2DM compared to FFDM alone (rather than the dual acquisition of FFDM+DBT). A range of different imaging units were used in these studies including Hologic Selenia Dimensions, Siemens Mammomat Inspiration and GE SenoClaire units. Findings from other studies that used other units may report different results. No systematic reviews assessing radiation dose for a range of different imaging modalities have been completed. Included studies are listed below. Methodologies for most included studies are reported in *section 3.1.1*. *Tables 20a-c* includes a summary of all study results.

Literature reviews

One review: Rocha García & Fernández, 2019

Randomized controlled trials

One RCT (one paper): Aase et al., 2019

One randomized study embedded in population-based screening: Pattacini et al., 2018

Prospective trials embedded in population screening programs

Seven studies (eight papers): Houssami et al., 2019; Skaane et al., 2019; Bernardi et al., 2018; Caumo et al., 2018; Gennaro et al., 2018; Østerås et al., 2018; Romero Martín et al., 2018; Zackrisson et al., 2018

Summary of key findings published between 1 January 2018 and 31 December 2019

Radiation dose varies with the image technique process used (DBT+s2DM, FFDM alone or FFDM+DBT), the number of and type of views, the use of automatic exposure control, positioning, breast size and composition, and by DBT system used. Most of the literature regarding MGD published since December 2017 reported on MGD for DBT+s2DM compared to FFDM alone, although some new data on dual acquisition (FFDM+DBT) was also reported.

Mixed results were presented on MGD, including within imaging protocols. For DBT+s2DM compared to FFDM alone, most studies reported that MGD was higher when DBT+s2DM was used although the To-Be-1 RCT, the only study to utilise GE SenoClaire units, reported no significant difference in MGD between DBT+s2DM and FFDM alone. All the other studies reported increases with this DBT modality. Some increases were small (from about 30 percent more) but an increase in MGD of almost double was reported in the Maroondah pilot. The reasons for these differences are not entirely clear although some differences may be due to imaging systems or the use of automated dose (as opposed to radiologist setting). 'Real world' studies all reported a higher per-view MGD when DBT was used compared to FFDM alone (which is consistent with some of the previous technical evaluation findings). Further investigation into the use of different imaging units (Hologic Selenia Dimensions, Siemens Mammomat Inspirations and GE SenoClaire units), software (Volpara Solutions and Quantra™) and protocols in relation to the impact on radiation dose is needed as some studies have started to suggest potential inaccuracies within breast density and radiation dose estimation that could be impacting research results and potentially women being screened.

DBT+s2DM compared to FFDM alone

DBT+s2DM has been developed as a mechanism to reduce the higher radiation doses associated with dual acquisition (FFDM+DBT). Instead of separately acquiring the 2D image (which is important for the depiction of certain radiological features like microcalcifications), 2D images are generated from the DBT-acquired data. In theory, eliminating the need for 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 per DBT image acquisition). Since December 2017, several studies have reported on the 'real world' MGD associated with DBT+s2DM compared to FFDM alone, with mixed results presented.

RCT

In a sub-study (interim analysis) of the To-Be-1 RCT (Norway), Aase et al. (2019) included data about radiation dose for a subset of study participants (14,089/28,749 participants). About half of these women were in the experimental arm (DBT+s2DM), and half in the other arm (FFDM). DBT images (acquired using GE SenoClaire units with DBT enabled) were captured with nine low-dose exposures over an angle of 25°, which were then reconstructed into 1mm slices and 10mm slabs as well as synthesised into two-dimensional images. MGD per exposure was calculated from the raw image data and the MGD per screening examination was calculated as the sum of the radiation doses reported by the software for both views and breasts divided by two. No information about breast compression thickness was provided. Aase et al. reported no statistically significant difference between DBT+s2DM and FFDM in relation to MGD:

- DBT+s2DM: 2.96 mGy
- FFDM: 2.95 mGy (p=.433).

The To-Be-1 RCT was the only study to record using GE SenoClaire units for image acquisition, and one of two studies to report using Volpara Solutions software to determine MGD. In relation to the data capture system, Aase et al. reported that the manufacturer (GE) stated that the target MGD for DBT using automatic exposure control was equivalent to the MGD per view for DM. Aase et al. stated that the absence of a difference between MGD with DBT+s2DM and FFDM aligned with how the system is set to operate by the manufacturer. Hofvind et al. (2019), reporting the main analysis from this RCT, added that the radiation dose measured for DBT+s2DM in the To-Be-1 trial was lower than that reported in other studies and that this might have negatively affected the image quality; however, differences in vendor-specific technical implementation and the optimisation of mammography workstations can also affect image quality. These factors along with reader experience can also impact on CDR detection, and are factors to consider in reviewing the lack of statistical difference in CDR observed in the To-Be-1 RCT (see *section 3.1.1*).

Prospective trials embedded in population screening programs

New data on MGD from the Maroondah, Cordoba and Verona pilot evaluations and updated MGD data from STORM-2 were also reported (methodologies described in *section 3.1.1*). These 'real world' studies used Hologic Selenia Dimensions for DBT and FFDM alone imaging but sometimes used other units for FFDM imaging. All reported a higher per-view MGD when DBT was used compared to FFDM alone (which is consistent with previous technical evaluation findings).

A sub-study of the STORM-2 trial was reported by Bernardi et al. (2018). A single breast positioning was used to obtain each view (CC and MLO). No information about breast compression thickness was provided. The estimated MGD per view (all study participants) was higher for DBT (1.87 mGy, SD 0.67) compared to FFDM (1.36 mGy, SD 0.51). For FFDM+DBT, dual acquisition reported an MGD per view of 3.22 mGy (SD 1.16). The DBT+s2DM MGD reported by Bernardi et al. (2018) is similar to Strudley et al.'s 2014 practical evaluation data estimate (1.81 mGy).

A slightly higher MGD was reported in the Verona evaluation pilot (Caumo et al., 2018). The Verona Screening program used Selenia Dimensions units. Screening examinations included two-view (CC and MLO) DBT+s2DM of each breast compared to a historical cohort screened with FFDM. Using data from all participants, the MGD for a single DBT view was higher (2.09 mGy per view, SD 0.55, range 1.13–3.65 mGy) compared with single-view FFDM (1.48 mGy per view, SD 0.58, range 0.52-3.13 mGy per view). No information about breast compression thickness was provided.

Results from the Cordoba pilot (Romero Martín et al., 2018) were similar but MGD was much higher overall (i.e., MGD for DBT+s2DM was higher than for FFDM alone: 4.97 mGy compared to 3.27 mGy, breast compression thickness = 62.5mm). This may be because of participant characteristics as Romero Martín et al. only reported MGD from a subset of 149 randomly selected participants (no participant characteristics were provided).

In the Maroondah pilot study (Houssami et al., 2019), women receiving DBT+s2DM were screened with Hologic Selenia Dimensions 8000 unit. Women receiving FFDM were screened with either Hologic Selenia Dimensions 8000 or Siemens Mammomat Inspiration. MGD results from the Maroondah pilot were higher than the To-Be-1, STORM-2 or Verona results (acquisition was almost double the MGD compared to FFDM imaging). The MGD for DBT views captured on the Hologic unit were 2.55 mGy per view compared to 1.32 mGy per FFDM view captured on the Siemens unit and 1.42 mGy per view on the Hologic unit. Houssami et al. highlighted that data indicated that the breasts imaged in 3D mode were, on average, substantially thicker than those imaged in 2D mode. A possible explanation might be that less compression was used with DBT, but this is not supported by an examination of the compression data which is measured in Newtons (N). DBT+s2DM had a mean of 83 N [SD, 25 N] compared to 82 N [SD, 22 N] for FFDM. The authors thought it more likely that the accuracy of the compressed breast thickness indicated on the images and included in the Digital Imaging and Communications in Medicine (DICOM) header may have differed between the two units. Measurement checks undertaken suggest that the Hologic unit may overestimate thickness by as much as 3 mm, the Siemens unit by only 1 mm. This may partially explain the difference in thicknesses in the 2D and 3D data.

The significance of this finding is that it has implications for estimating MGD for a particular image because the internal dose model assumes that the indicated breast thickness is the true thickness. Using the displayed MGD value as an indicator of relative dose may therefore be problematic when more than one machine is used. As the Hologic DBT MGD values were higher than those of Siemens digital mammography (mean ratio: 1.9), Houssami et al. suggested that these results be considered with caution due to the different image acquisition units employed in the study. More research in this area is needed to determine if the differences seen in mGy between DBT+s2DM

and FFDM are in part related to the image acquisition units employed, compression levels, software or some other factor.

FFDM+DBT compared to FFDM alone

The following results are reported here for completeness, rather than reflecting current screening practices where DBT has been trialled or implemented. In these papers, per-view MGD for DBT was higher than for digital mammography.

Literature review

In their narrative literature review, Rocha García & Fernández (2019) compared the MGD between different modes of acquisition of FFDM+DBT and DBT alone (including looking at papers covered in *Allen + Clarke*'s previous literature reviews) as well as papers by Michell et al. (2012) and Feng et al. (2012). Rocha García & Fernández reported that use of DBT alone resulted in a 39 to 48 percent radiation dose reduction compared to FFDM+DBT; however, caution is needed with these results as Michell et al.'s study included women recalled from a screening program (and therefore undergoing a different suite of imaging) and Feng et al.'s study used breast phantoms to investigate FFDM and DBT radiation doses.

Randomized controlled trial

Pattacini et al. (2018) reported baseline data from the Reggio Emilia RCT (methodology described in *section 3.1.1*). Data on MGD was collected from all study participants and was generally very high (more aligned to the Cordoba trial results than other studies). MGD results for FFDM were 4.84 mGy (IQR, 4.24–5.72) and for FFDM+ DBT = 6.40 (IQR, 5.68.–7.36). The dose in the experimental arm was 2.3 times higher than in the non-experimental arm. The imaging unit used was GE Senographe Essential digital systems (GE Healthcare, Buc, France). The authors make no comment on why the MGD values were so high.

Prospective trials embedded in population screening programs

In the OTS trial conducted in Norway (Østerås et al., 2018) (methodology described in *section 3.1.1*), the authors analysed paired same-compression FFDM and DBT acquisitions (n=3,819), using the Dance model they determined radiation dose and reported directly from the DICOM metadata. Østerås et al. reported the following results per view:

- FFDM MGD: 1.72 mGy
- DBT MGD: 2.09 mGy, and
- FFDM+DBT MGD: 3.81 mGy[29].

The mean compressed breast thickness in this study was 53.4 mm (SD 12.9, range 14–101) for all screens. The mean dose ratio between FFDM and DBT was 1.24 mGy. Denser breasts (BIRADS

[29] Skaane et al. (2019) reported on the overall MGD for the OTS study population, stating the following (which are lower than the sub-study population rates reported by Østerås et al. (2019):
- FFDM MGD: 1.58 mGy
- DBT MGD: 1.95 mGy, and
- FFDM+DBT MGD: 3.53 mGy.

categories 3 and 4) generally had a lower MGD (FFDM mean average glandular dose (AGD) was 1.73 compared with 1.79 mGy for DBT) than fatty breasts (FFDM mean AGD was 1.74 compared with 2.27 mGy for DBT,). Østerås et al. suggested that as denser breasts generally had a lower DBT/FFDM AGD ratio than fatty breasts, this implied a lower 'dose penalty' for using DBT on dense breasts compared to non-dense breasts.

The study also undertook an investigation into differences in mean AGD between estimates based on measurements (air kerma and HVL) versus data derived from DICOM header data and found a difference of 3.8 percent, but for one mammography unit as high as 7.9 percent. Mean error of using the AGD value reported in the DICOM header was 10.7 percent and 13.3 percent respectively. They concluded that measurement of breast density, radiation dose and beam quality can substantially affect AGD estimates. This provides a cautionary note for further research to consider this potential variation.

These figures differ slightly from those presented by Skaane et al. (2019) for the same trial, who noted that the average glandular dose for FFDM was 1.58mGy (±0.61 mGy) compared to 1.95 mGy (±0.58 mGy) for DBT alone, and 3.53 mGy (±0.84 mGy). This may be due to the different sub-sets of data used (i.e., Skaane et al. presented data for all study participants).

DBT_{MLO} compared to other imaging combinations

Two studies reported on the MGD for single projections, with one study using a wide-angle, reduced compression reporting a lower MGD with DBT_{MLO} (the Malmö trial) and one reporting a higher MGD (results from a STORM-2 sub-study) (methodologies described in *section 3.1.1*).

The Malmö Breast Tomosynthesis Screening Trial compared wide-angled DBT_{MLO} using Siemens Mammomat Inspirations to FFDM and explored the value of adding DM_{CC} to DBT_{MLO}. Zackrisson et al. (2018) reported that the MGD was lower in DBT_{MLO} (2·3 mGy, SD 0·7) compared with FFDM (2·7 mGy, SD 0·8). The mean compression force for DBT_{MLO} was lower (71 N; SD 21) compared to FFDM (118N; SD 24), a mean reduction of 40 percent in the compression force. This study demonstrated that breast cancer screening by use of DBT_{MLO} with a reduced compression force has higher sensitivity at a slightly lower specificity for breast cancer detection compared with FFDM and has the potential to reduce the radiation dose, participant discomfort and screen-reading burden.

Another STORM-2 sub-study was completed by Gennaro et al. (2018). We have included this because it compared MGD for single projections (including DBT_{MLO}). The authors used a subset of 1208 randomly selected screening patients, which resulted in 4768 paired CC and MLO views of DBT and FFDM images acquired using a Selenia Dimensions system (Hologic) unit. Fifteen DBT projections within 15° were captured using the automatic exposure mode. The results indicated statistically significantly higher MGD for both DBT projections compared to DM projections:

- DBT_{CC} was 1.858 mGy; DM_{CC} was 1.366 mGy (p<.0001), and
- DBT_{MLO} was 1.877 mGy; DM_{MLO} was 1.374 mGy (p<.0001).

BlandAltman analysis showed that the average increase of DBT dose compared to FFDM was 38 percent (0-75). MGD increases with breast thickness, for both FFDM and DBT but decreases with

lower breast density. For a given breast thickness or density, MGD is generally higher with DBT than with FFDM. MGD is higher for lower compression force and compression pressure, in both FFDM and DBT. Finally, MGD is proportional to both x-ray beamHVL and mAs level; however, the automatic exposure control operates differently in FFDM and DBT; therefore, the MGD increase with mAs in FFDM is approximately linear, while dose increase is faster in DBT because it is carried out using more penetrating spectra (heavier filtration/ higher kVp).

Table 20a: DBT+s2DM compared to FFDM+DBT or FFDM alone or DBT alone: studies reporting on radiation dose (MGD per view, mGy)

Study	Sample	Study type	DBT+s2DM MGD per view mGy (SD)	FFDM+DBT MGD per view mGy (SD)	FFDM alone MGD per view mGy (SD)	Additional information
RCT						
Aase et al. (2019) To-Be-1 (sub-study)	14,089 women randomized to DBT (n=7155) or FFDM (n = 7119)	Parallel group RCT using GE SenoClair, using Automatic Exposure Control (AEC), MGD determined using Volpara Solutions.	2.96	NA	2.95 (*p*=.433)	MGD = Sum of the radiation doses for both views and breasts divided by two.
Prospective trials embedded in population screening programs						
Houssami et al. (2019) Maroondah	10,146 women DBT+s2DM: 4993 women (5018 exams) FFDM alone: 5153 women (5166 exams)	Prospective pilot trial using Hologic Selenia Dimensions 8000 (+/- C- View™) or Siemens Mammomat Inspiration DBT – 4 views DBT and FFDM images acquired with a single compression per view.	3D Hologic unit < 36 mm = 1.12 36–45mm = 1.34 46–55mm = 1.69 56–65mm = 2.21 66–75mm = 2.88 >75mm = 3.67 All = 2.55	NA	2D Siemens unit < 36 mm = 0.93 36–45mm = 1.01 46–55mm = 1.15 56–65mm = 1.30 66–75mm = 1.50 >75mm = 1.92 All = 1.32	Ratio < 36 mm = 1.21 36–45mm = 1.33 46–55mm = 1.47 56–65mm = 1.70 66–75mm = 1.91 >75mm = 1.92 All = 1.92
Caumo et al. (2018) Verona	DBT+s2DM: 16,666 and FFDM: 14,423 women previously screened in the Verona program	Prospective pilot evaluation using Hologic Selenia Dimensions	2.09 (0.55)	NA	1.48 (SD 0.58)	NA
Romero Martín et al. (2018) Cordoba	16,067 women in FFDM+DBT with s2DM images 3341 = prevalent 12,727 = incident	Prospective transversal study using Hologic Dimensions system	4.97 (±1.28)	8.24	3.27 (±0.83)	Mean CBT 62.5 mm (12.8)

Table 20b: FFDM+DBT compared to FFDM alone: studies reporting on radiation dose (MGD per view, mGy)

Study	Sample	Study type	FFDM+DBT MGD per view mGy (SD)	FFDM alone MGD per view mGy (SD)	DBT alone MGD per view mGy (SD)	Additional information
RCT						
Pattacini et al. (2018) Reggio Emilia	Women aged 45-74y attending for incident screening (women aged 45-49y screened annually; women aged ≥50 screened biennially). Women with family history were excluded. FFDM: 9783 women (mean age 56.3y) FFDM+DBT: 9777 women (mean age 56.3y)	Baseline data from a prospective randomized controlled trial using a test and treat methodology using GE Senographe Essential systems. (No AEC information available)	FFDM+DBT = 6.40 (IQR, 5.68, 7.36)	FFDM = 4.84 (IQR, 4.24, 5.72)	NA	The dose in the experimental arm was 2.3 times higher than. 49 women who had problems completing DBT because of compression discomfort.
Prospective trials embedded in population screening programs						
Bernardi et al. (2018) STORM	9672 women	Prospective trial with paired data using Selenia® Dimensions Unit operated in COMBO© mode; Hologic and using C-View™	1.87 (0.67)	3.22	1.36 (0.51)	NA
Østerås et al. (2018) OTS	3819 women mean age of 58.8 with a range of 50–70 years resulting in 15276 paired DM and DBT views. CC and MLO views were acquired of both breasts in a combo mode (DM plus DBT acquisition during the same compression).	Cohort study drawn from OTS using Hologic Selenia Dimensions units, using automatic exposure control setting	All: 3.84 Non-dense: 4.01 Dense: 3.51 (p<.001).	All: 1.74 Non-dense: 1.74 Dense: 1.73 (p=.13)	All: 2.10 Non-dense: 2.27 Dense; 1.79 (p<.001).	DBT/FFDM ratio All: 1.24 Fatty: 1.33 Dense: 1.08 (p<.001).
Skaane et al. (2019) OTS	24,301 women screened in the Norway BreastScreen program	Fully paired trial using Hologic units	3.53 (0.84)	1.58 (0.61)	1.95 (0.58)	NA

Table 20c: DBT_{MLO} compared to other imaging combinations: studies reporting on radiation dose (MGD per view, mGy)

Study	Sample	Study type	DBT_{MLO} /cc MGD per view in mGy (SD)	FFDM alone MGD per view in mGy (SD)	Additional information
Prospective trials embedded in population screening programs					
Gennaro et al. (2018) STORM-2	A subset of 1208 screening patients was randomly taken from the STORM-2 trial population (4,768 paired views of DBT and FFDM images)	Prospective RCT using paired data using Selenia Dimensions system (Hologic); for the DBT mode, 15 projections were acquired within 15° Automatic exposure mode. MGD determined using Volpara v. 1.5.2.0.	CC =1.858 MLO = 1.877	CC = 1.366 (p=.0001) MLO = 1.374 (p<.0001)	FFDM and DBT images were acquired in COMBO mode, i.e. with the same breast positioning and compression pressure. MGD is higher for lower compression force and compression pressure[30], in both FFDM and DBT.
Zackrisson et al. (2018) Malmö trial	14,848 women aged 40-76 years (mean age = 57 years) presenting for screening in Malmö, Sweden	Prospective paired trial where women underwent FFDM and wide-angled one-view DBT_{MLO} using Siemens Mammomat Inspirations.	MLO = 2·3 (0·7)	MLO = 2·7 (0·8)	NA

[30] Compression pressure; pressure is calculated by compression force divided by the breast contact area.

4. IMPLEMENTATION OF DBT AS A SCREENING TOOL

Much of the published literature focuses on DBT's sensitivity, specificity and safety with the studies reporting on the nature of the association between DBT (either alone, as an adjunct to FFDM, or with s2DM) and specific clinical outcomes and short-term performance metrics. There is growing confidence that as a screening strategy DBT could enhance a screening program with several programs moving to pilot DBT as the screening test (eg, the Trento and Verona screening programs). Another key area of research to consider is implementation. Data from the studies that have implemented DBT are now reporting on the 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, and
- cost.

4.1. Image acquisition

An increase in image acquisition time could increase the discomfort felt by women when participating in breast screening. Additionally, an increase in time could impact high-throughput clinic workflow.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

Reports from practical evaluations noted that, for DBT imaging, 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, use of DBT is usually associated with only a small amount of extra time investment for women and radiographers. In the STORM trial, image acquisition was defined from a sub-set of 20 screening examinations (from the start of the first-view breast positioning to compression release at the last view). It took longer to acquire images with FFDM+DBT compared to FFDM alone:

- FFDM+DBT: 4 min 3 sec (range = 3 min 53 sec to 4 min 18 sec), and
- FFDM alone: 3 min 13 sec (range 3 min 0 sec to 3 min 26 sec).

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 sec per view to obtain (an additional 40 sec overall with the use of DBT as an adjunct screen; 3 min 55 sec).

Updated findings

Literature published since 31 December 2017 on image acquisition time is limited (only results from the To-Be-1 RCT). The methodology for the To-Be-1 trial is described in *section 3.1.1.* No data on image acquisition was reported from the Maroondah pilot (Houssami et al., 2019). We are mindful that many of the other pilot evaluations may have recorded information about image acquisition but not yet reported on this or on the impact on workflow that an additional time to acquire images might have. This may be an area where further research is published in the future, or it may be covered in some of the other studies currently underway (see *section 1.5*). No systematic reviews assessing image acquisition for a range of different imaging modalities have been completed. Included studies are listed below.

Literature review

One narrative literature review: Rocha García & Fernández, 2019

Randomized controlled trial

One RCT (two papers): Aase et al., 2019; Moger et al., 2019

Summary of key findings published between 1 January 2018 and 31 December 2019

An increase in image acquisition time could increase the discomfort felt by women when participating in breast screening. Additionally, an increase in time could impact high-throughput clinic workflow. Currently available literature is surprisingly sparse, and certainly insufficient to determine the impact of DBT on image acquisition time. Only one primary study (the To-Be-1 RCT) reported findings on image acquisition time. Results from this RCT were based on the time the woman entered the exam room until the time she left. Results reported from the To-Be-1 RCT stated that DBT+s2DM took 54 sec longer than FFDM (DBT+s2DM imaging took 5 min 24 sec compared to four min 19 sec with FFDM). While the definition of image acquisition differed to other previous studies (being the time in the exam room, rather than the machine settings), the trend appears to remain the same: DBT image acquisition takes longer than FFDM. One explanation presented by the To-Be-1 RCT is that additional time for DBT may have been impacted by explaining to the woman the new technology. We do not know if the actual image acquisition time was much longer (one second is not much more, as reported in other studies). There is also a suggestion that newer machines have the potential to reduce the difference in DBT capture time (observed in early commentary from Moger et al. (2019) in relation to the extension the To-Be-1 RCT (To-Be-2). The Maroondah pilot in Australia may also provide more insight into these issues in the future but it has not yet reported any data on image acquisition.

DBT+s2DM compared to FFDM+DBT and/or FFDM alone

Literature review

Rocha García & Fernández (2019) reported on an earlier Bernardi et al. (2012) study, which stated that image acquisition takes approximately four sec per projection for FFDM, while a dual acquisition of DBT and DM takes 26 percent longer (approximately five seconds per projection);

however, this used a different measure from the To-Be-1 RCT (just the machine use time, not the women's time in an examination room).

RCT

Reporting on a sub-study (interim analysis) for the first year of the To-Be-1 RCT, Aase et al. (2019), defined 'examination time' as the time spent from when the woman entered the examination room until she left. This was manually registered using a stopwatch for randomly selected women screened with DBT+s2DM (n=438) and FFDM (n=535) during March 2017. Two views (CC and MLO) were obtained of each breast with single breast compression per view. Women spent an average time of five min 24 sec (median = 5 min 13 sec) for DBT and 4 min 19 sec (median = 4 min 7 sec) for FFDM in the screening examination room (p<.01).

In the main analysis of this RCT, Hofvind et al. (2019) proposed that the increased examination time was mainly due to additional time spent on explaining to the women how the x-ray machine would move and to make the x-ray tube ready for exposure. No information about the actual machine acquisition time was reported in any of the papers on the To-Be-1 data published to date. This extra time is expected to be reduced or resolved in subsequent screening rounds.

Moger et al. (2019) discussed preliminary results from To-Be-2 study (the five-year follow-up study to To-Be-1) where women are being screened with the newer GE Senographe Pristina machine. The authors indicated potential for shorter examination times for DBT compared to FFDM than the one-minute increase observed in the first RCT, which would align time in the examination room to that seen when FFDM is used. This has significant implications for clinic workflow; however, further information is needed to confirm this change.

4.2. Reader performance

Reader performance is a critical component of any population-based breast screening program. The ability of those who are responsible for accurately interpreting breast screen images (primarily radiologists) to determine true positive (sensitivity) and true negative (specificity) impacts for the individual woman, the screening unit and the population-based screening program.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

Few studies reported stratified reader sensitivity and specificity data in relation to experience levels and reading strategy. Most studies involved readers with a range of experience in breast screening and radiology in general. Consistent evidence (mostly based on small numbers of cases and readers but including a large sub-study from the TOMMY trial) indicated that less experienced readers improved more when using either FFDM+DBT or $DBT_{MLO/CC}$ compared to FFDM, but improvements were noted for everyone once a learning curve had been passed. It is unclear whether this improvement reflected the development of less experienced practitioners' competence, or whether DBT is 'easier' to read without as much experience in breast cancer imaging.

Inter-observer agreement is an important measure of the overall accuracy of data collected to detect and evaluate breast lesions. Based on smaller retrospective observer studies in a range of diagnostic groups, evidence suggested an increase in inter-observer agreement for BIRADS classification with the use of DBT compared to DM. This increase in agreement was observed in the following DBT reading protocols: FFDM+DBT compared to FFDM alone, DBT alone compared to FFDM alone, or s2DM. Inter-observer agreement (as measured by kappa statistic) with DBT imaging increased in all studies that reported it, with the use of DBT increasing overall agreement from moderate to good or excellent. Reported increases were usually about 25 percent. FFDM+DBT appears to be a more reliable test for BIRADS agreement with kappa statistics exceeding 0.8 with much lower rates reported for FFDM (ranging from 0.58 to 0.87). The use of s2DM compared to FFDM also appears to improve inter-observer agreement, suggesting that lesion conspicuity is greater with the reduction in 'noise' available with the 3D reconstruction algorithm.

Updated findings

A small amount of literature published since 31 December 2017 reported results for reader performance including results from a number of studies embedded in population-based screening programs. The findings cover stratified reader sensitivity and specificity data in relation different imaging methods and experience in addition to consideration of factors that help or hinder reader performance such as training in DBT interpretation, availability of prior DBT images and technologies. No systematic reviews assessing image acquisition for a range of different imaging modalities have been completed. Included studies are listed below.

Literature reviews

Two literature reviews: Chong et al., 2019; Rocha García & Fernández, 2019

Randomized controlled trial

One RCT (two papers): Aase et al., 2019; Hofvind et al., 2019

Prospective studies embedded in population screening programs

Six studies: Houssami et al., 2019; Bernardi et al., 2018; Caumo et al., 2018; Hofvind et al., 2018; Romero Martín et al., 2018; Zackrisson et al., 2018

Observational studies

One study with a prospective design: Miglioretti et al., 2019

Two studies with a retrospective design: Chae et al., 2019; Simon et al., 2019

Summary of key findings published between 1 January 2018 and 31 December 2019

Few studies described performance metrics related to reader performance and / or undertook stratified analysis of performance across individual radiologists or between groups of radiologists with differing levels of experience (although most studies involved readers with a range of experience in breast imaging and some noted the level of experience in interpreting

DBT and s2DM images). Some studies noted pre-trial trial training in DBT and s2DM screen reading; however, most studies also recognised that this was either insufficient or as in the case of the Malmö trial, not sufficient in a 'real world' setting. It is difficult to determine the extent to which training and prior experience made a difference across the studies that mentioned it. Many studies noted that there is a likely presence of a learning curve, meaning that recall rates tend to drop over time as readers become more familiar with the different type of image that DBT and DBT+s2DM provides. Some studies reported that the learning curve was very short on average with sustained improved screening performance soon after DBT adoption; however, changes in performance with experience may vary across radiologists.

The STORM-2 trial was one a few studies that provided individual performance data, which indicated that while cancer detection rates across the team of radiologists improved with the addition of DBT, so did the false positive recall rate. Availability of prior DBT images was also considered to be influential to reader performance.

Further investigation is warranted to determine if the experience and training in DBT / s2DM screen reading and the availability of prior images is significantly impacting interpretation times, recall rates, false positive recall rates and cancer detection rates.

Literature reviews

Two narrative literature reviews discussed reader experience and performance: Rocha García & Fernández (2019) and Chong et al. (2019). Neither of these literature reviews discussed one specific imaging modality alone, instead they provided overarching views on reader performance. Rocha García & Fernández commented that there is a learning curve for both the technician and the radiologist who interprets the DBT images that is similar to training to use any new application or equipment. They cited the FDA's Mammography Quality Standards Act and Program, which requires staff to receive eight hours of initial training before independently using any new form of mammography. Rocha García & Fernández suggested that the number of recalls may increase initially (representing a learning curve), but usually stabilises at six months and then starts to fall. We explore this understanding further in the primary studies below.

Chong et al. (2019) commented briefly on reader confidence and efficiency in interpreting DBT images. The authors noted that while reader confidence in image interpretation is likely improved by the clarity of DBT images, the additional time required to read the images is important as well (see *section 4.3*). Chong et al. also noted advances in presentation modes that might increase reader efficiency such as automated "slabbing" and machine learning–based detection algorithms as well as machine-learning CAD.

DBT+s2DM compared to FFDM+DBT and/or FFDM alone

RCT

A pool of eight breast radiologists undertook the initial screen readings in the To-Be-1 Norwegian trial (Hofvind et al., 2019). A stringent consensus process was also used, whereby if either radiologist assigned a score of BIRADS two or higher, consensus was used to determine whether to recall the woman for further assessment, which was done by pairs of radiologists. A third radiologist was consulted if the pair could not agree. The experience of radiologists in screen reading (screen film and digital mammography) before start-up of the trial varied from zero to

approximately 110,000 examinations. The program's recommendations of 5000 annual screen readings was met by half of the radiologists. Hofvind et al. reported that all participating radiologists were trained in DBT screen reading and diagnostics to some extent before the start of the trial although it is not clear whether this was completely consistent as data from Aase et al.'s 2019 sub-study analysis also reported that not all To-Be-1 RCT radiologists participated in a screen reading DBT pilot eight weeks before the trial commenced (which involved reviewing about 300 DBT screening examinations). Hofvind et al. reported that radiologists did not have the opportunity to practice DBT screen reading in an everyday screening setting until the RCT commenced. This is consistent with Aase et al.'s suggestion that the first year of the To-Be-1 RCT be considered a learning period.

Prospective trials embedded in population screening programs

For studies that investigated CDR by reader, all demonstrated variation between readers with some increasing CDR and others not. Romero Martín et al. (2018) noted statistically significant improvements for three out of five readers and with one reader showing no difference in performance.

In a sub-study of the STORM trial, Bernardi et al. (2018) reported on the association between incremental CDR and radiologist experience in screen-reading mammography. This was assessed across the seven breast radiologists who participated in the trial. Participating radiologists had an average of 13 years' experience in breast imaging (range = 3–23). The radiologists each received training in DBT and had been using it for an average of 2.7 years (range = 2 to 3 years) at trial initiation. The findings from Bernardi et al. (2018) are presented in detail as they are one of few studies that provide insight into individual reader performance. With respect to FFDM with and without DBT, true positive detection (relative sensitivity) in FFDM ranged from 46 percent to 100 percent (median = 59.5 percent). Sensitivity increased with FFDM+DBT, ranging from 75 percent to 100 percent (median = 76 percent). For all but one radiologist, screen-reading with FFDM+DBT improved breast cancer detection over FFDM alone.

With respect to s2DM with and without DBT, true positive detection (relative sensitivity) ranged from 56 percent to 76 percent (median = 64 percent) for s2DM. Relative sensitivity for DBT+s2DM ranged from 67 percent to 88 percent (median = 78 percent). For all radiologists, screen-reading with DBT+s2DM improved breast cancer detection over s2DM alone.

Table 21a-b provide details of the STORM-2 radiologist-specific cancer detection measures at screen-reading using FFDM+DBT and DBT+s2DM. The number of false positive recalls for most radiologists was modestly higher for FFDM+DBT than s2DM alone.

Table 21a: STORM-2 radiologist-specific CDR using FFDM and DBT

FFDM alone Readers	FFDM+DBT Readers	FFDM+DBT Additional cancers detected by integrating methods
1= 100% (5/5) 2= 57% (16/28) 3= 58% (25/43) 4= 61% (27/44) 5= 46% (24/52) 6= 65% (13/20)	1= 100% (5/5) 2= 75% (21/28) 3= 77% (33/43) 4= 75% (33/44) 5= 83% (43/52) 6= 75% (15/20)	1= 0% (0/5) 2= 18% (5/28) 3= 19% (8/43) 4= 14% (6/44) 5= 37% (19/52) 6= 10% (2/20)

Table 21b: STORM-2 radiologist-specific sensitivity measures at screen-reading using s2DM and DBT

S2DM Readers	DBT+s2DM Readers	S2DM+DBT Additional cancers detected by integrating methods
1= 56% (5/9) 2= 57% (21/37) 3= 64% (23/36) 4= 70% (26/37) 5= 76% (25/33) 6= 69% (11/16) 7= 63% (15-24)	1= 78% (7/9) 2= 76% (28/37) 3= 86% (31/36) 4= 86% (32/37) 5= 88% (29/33) 6= 75% (12/16) 7= 67% (16/24)	1= 22% (2/9) 2= 19% (7/37) 3= 22% (8/36) 4= 16% (6/37) 5= 12% (4/33) 6= 6% (1/16) 7= 4% (1/24)

While this study noted variability in the magnitude of effect from integrating DBT on individual radiologist's true positive and false positive detection, there was an overall pattern of increasing cancer detection and also increasing false positive recall for most readers. The study concluded that there was no statistical evidence between incremental CDR and radiologist experience in screen-reading mammography, and the rank correlations were consistent with these results. Additionally, as the STORM-2 trial represented the radiologists' first screening experience using s2DM images, the authors propose that increasing experience with s2DM will reduce FP's.

The BreastScreen Norway trial (Hofvind et al., 2018) and the Verona trial (Caumo et al., 2018) noted that radiologists had limited experience with DBT/s2DM. This inexperience was often cited as a limitation or potential study effects. Each of these studies reported on different aspects of reader performance. For example, Caumo et al. (2018) noted a very high CDR (9.3 per 1000 screening examinations): Verona was a study site for STORM-2 and this may have resulted in increased experience for all readers (which positively influenced CDR). The BreastScreen Norway trial (Hofvind et al., 2018) noted the breast imaging experience level of the interpreting radiologists was generally high; however, experience with interpretation of DBT and s2DM images varied. The majority of radiologists in the FFDM group performed screen reading before and during the study period. This was not the case in Oslo, where most radiologists had limited or no experience with DBT+s2DM. The authors made no comment on overall reader performance but noted that the limited experience in use of DBT could increase interpretation time, increase recall rate, and decrease CDR (although these results were not observed).

The Maroondah pilot trial (Houssami et al., 2019) used independent double reading with independent arbitration to investigate the detection measures for FFDM compared to DBT+s2DM. The seven radiologists who assessed 10,146 women's screening examinations each had approximately five years' clinical experience in using DBT to assess mammography-detected findings and had received additional training in DBT screen reading before the trial commenced. While Houssami et al. did not (at this stage) provide analysis of reader performance, the authors proposed that the longer screen reading times (see *section 4.3*) and higher recall rate (see *section 3.5.1*) seen with DBT+s2DM may decline in Australia as the 'novelty factor disappears' and screen readers become more experienced with the technique and have access to DBT screens from earlier screen rounds that can be compared with current screens.

Retrospective analysis

In a small retrospective observational single site study conducted in the US, Simon et al. (2019) undertook ROC analysis to evaluate the ability of readers to differentiate patients with and without cancer using a likelihood of malignancy scale (0–100%). The enriched study sample consisted of images from 89 women with confirmed cancer and 100 women without cancer (mean clinical follow-up was 2.6 years). The images were read by three independent radiologists with 2.5 years' experience reading DBT+s2DM images, 2.5 to 6 years of reading DBT+FFDM images and 3.5 to 16.5 years' experience in reading FFDM images. The ROC analysis was performed for each reader and for the averaged response of all three readers. Interrater agreement on the likelihood of malignancy was substantial. The intraclass correlation coefficient for DBT+s2DM was 0.62 (95%CI: 0.55, 0.69), and the intraclass correlation coefficient for FFDM+DBT was 0.69 (95%CI: 0.62, 0.75). There was no statistically significant difference between the AUC of DBT+s2DM and the AUC of FFDM+DBT for any reader (p=.1637) or for all readers averaged (p=.7888). While the study has numerous limitations (most notably its small, enriched sample), it provided relevant information regarding consistency in reader performance for three experienced radiologists. Its findings are similar to those of previous studies (i.e., use of DBT does not result in inferior performance by readers). Additionally, no significant difference was found in relation to differences in availability of prior screens (which has been cited by other studies). More larger scale studies with more robust methodology would enhance understanding of the impact of reader experience and image availability.

DBT alone compared to FFDM alone

Prospective analysis

A multisite (n=53) prospective study of Breast Cancer Surveillance Consortium data (271,362 participants) reported by Miglioretti et al. (2019) evaluated screening DBT performance by cumulative DBT volume within two years of DBT adoption relative to FFDM performance one year before adoption. The study analysed within-radiologist changes (breast imaging subspecialists compared to non-breast imaging subspecialists) in DBT performance with increasing cumulative interpretive volume of screening and diagnostic DBT across a large set of radiologists (n=104). The study highlighted that DBT unmasks many more small imaging findings that radiologists must learn to recognise as benign and subtle areas of architectural distortion that they must learn to recognize as suspicious. The authors found that the learning curve for acquiring DBT interpretive skills is very short, on average, with sustained improved screening performance soon after DBT adoption; however, changes in performance with experience may vary across radiologists. A summary of the study's findings is listed below:

- Recall rates over time: the mean recall rate was higher for screening FFDM one year before DBT adoption (10.4 percent; 95%CI: 9.5, 11.4) than for DBT (9.4 percent; 95%CI: 8.2, 10.6) (p=.02). This suggests that confidence may grow over time with more exposure to DBT images.
- Recall rates between reader groups: the difference in recall rates for FFDM versus DBT was not smaller for breast imaging specialists (10.0 percent compared to 9.7 percent, p=.60) than for readers who were not breast imaging subspecialists (11.0 percent

compared to 9.6 percent, p=.02). This suggests that those who do not specialise in breast screening may require additional training in DBT.

- Relative to FFDM, the recall rate for the lowest cumulative volume subgroup decreased within radiologists (OR=0.83; 95%CI: 0.78, 0.89; p<.001) and continued to remain lower for all volume subgroups (p<.001 for all subgroups). This suggested improved performance with DBT regardless of how much reading is being undertaken.

Overall, improvements in performance were similar for both reader groups and were sustained with increasing DBT volume in both groups. This suggested that, once trained in DBT, improvements in performance should be maintained, therefore targeted upfront training is worth the investment.

DBT compared to other imaging combinations

Prospective trials embedded in population screening programs

The Malmö trial (Zackrisson et al., 2018) compared the difference in reader sensitivity and specificity for single view DBT$_{MLO}$ compared with FFDM. Seven radiologists were involved in the trial and had a range of two to 41 years of experience in breast radiology (five readers had more than 10 years). All readers gained experience in DBT as part of their clinical work or in previous studies. No prior DBT images were available to the radiologists, but if prior FFDM images for women were available they were reviewed by the radiologist as part of the reading protocol for both arms of the study. The number of DBT false positive cases was found to decrease with increasing reader experience as depicted in *Figure 2* (below).

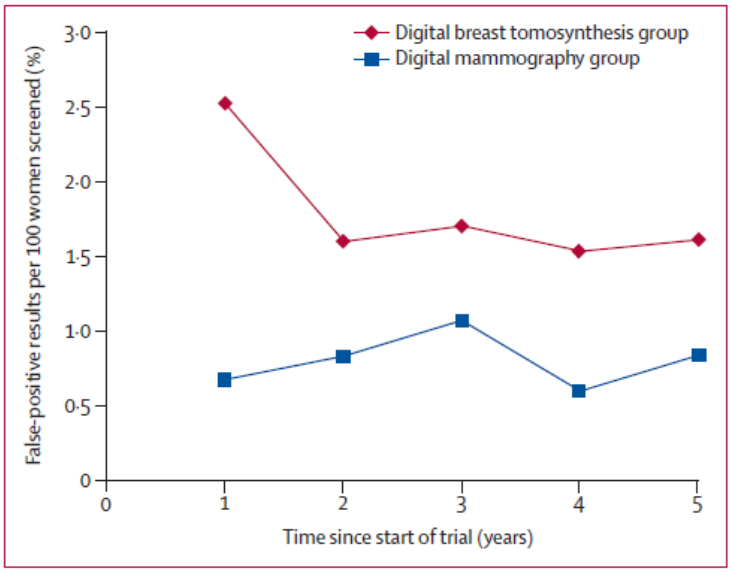


Figure 2: Number of false-positive results among recalled participants over the trial period for the two reading groups

Observational studies: retrospective design

In a retrospective observational study, Chae et al. (2019) reported on the diagnostic performance and interpretation time of DBT for both novice and experienced readers with and without using a computer-aided detection (CAD) system was investigated. We have included this study because

it is one of the few that look at the use of CAD in the interpretation of results. The role of CAD may be important in terms of reducing reading volume/time (see *section 4.3*). In Chae et al.'s study, image interpretation was performed by four radiologists (two novice readers who had one year of breast imaging experience and two dedicated breast radiologists with 14 and six years' experience). Each reader completed two reading sessions: with and without CAD and were blinded to readings of other radiologists, clinical information, and histological diagnosis. Three different scales were used to record each reader's interpretation: Modified BIRADS, probability of malignancy scale (a 1 to 7-point scale where 1 was definitely not cancer and 7 was definitely cancer), and percentage probability of malignancy scale. Scales were used to calculate diagnostic performance and ROC. For the BIRADS scale, the average AUCs across readers were 0.778 with CAD (95%CI: 0.733, 0.822) and 0.776 without CAD (95%CI: 0.732, 0.821). The difference in AUC with and without CAD was not statistically significant, and the results were consistent with the results obtained on using the probability of malignancy and percentage probability of malignancy scales. Although there were variations in the AUC values between the novice and experienced readers, results obtained with CAD were consistent in all readers. Differences in sensitivity and specificity were not statistically significant in the presence or absence of CAD system. Results were consistent across three different rating scales, additionally novice readers performed as well as the experienced readers in all aspects with and without CAD. This may have been influenced by the small and enriched sample of positive cancer cases.

4.3. Interpretation time

Any significant increase in interpretation time for DBT images compared to traditional FFDM will have an effect to screening clinic workflow, resources and costs. While some proponents of DBT state the additional time required for DBT is justified by increased CDR and PPV, any additional costs in a population screening program needs to be carefully considered.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

The literature reported a strong, consistent theme that implementation of DBT as an adjunct to FFDM increases reading and interpretation time. DBT produces many more images than FFDM alone (up to 25 per projection compared to two). As such, readers must 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, radiologist/reader resourcing and cost implications if DBT is implemented into population screening programs.

Updated findings

Two literature reviews, one RCT and three prospective studies presented findings relating to image interpretation time comparing DBT+s2DM with FFDM and FFDM+DBT with FFDM. No systematic reviews assessing image acquisition for a range of different imaging modalities have been completed. Methodologies for most included studies are reported in *section 3.1.1*. Included studies are listed below. A summary of interpretation times by study is included in *Table 23*.

Literature reviews

Two reviews: Chong et al., 2019; Rocha García & Fernández, 2019

Randomized controlled trials

One RCT (three papers): Aase et al., 2019; Hofvind et al., 2019; Moger et al., 2019

One randomized study embedded in a population screening program: Pattacini et al., 2018

Prospective trials embedded in population screening programs

Four studies: Houssami et al., 2019; Bernardi et al., 2018; Caumo et al., 2018; Romero Martín et al., 2018

Observational studies

One retrospective analysis: Chae et al., 2019

Summary of key findings published between 1 January 2018 and 31 December 2019

Studies observed increases in interpretation time for all DBT imaging, with a doubling of reading time being the most common (which is similar to earlier reported increases).

For DBT+s2DM compared to FFDM, the To-Be-1 RCT reported an initial reading time of 1 min 6 sec with DBT+s2DM compared to 39 sec with FFDM. The same was found in the Reggio Emilia RCT: interpretation time for FFDM+DBT was 56 sec compared with 34 sec for FFDM alone. Only one study reported almost no difference in interpretation time between DBT alone and DBT+FFDM if the result was positive; however, no specific data was provided so it is not possible to verify this. Increased reading time will have workflow and reader/radiologist resourcing implications. Some authors suggested that this additional time may be acceptable if DBT delivers less need for consensus (the number of screens requiring arbitration), lower false positive recall rates and higher CDR rates. There continues to be a need for more research into these variables and also with regard to the impact of DBT training and experience on interpretation time.

There was emerging evidence that the addition of computer-aided detection (CAD) to DBT may reduce interpretation time compared with DBT alone; however, this was only explored in one study and there was no comparison to FFDM. One small cancer enriched study provided positive results for the addition of CAD to DBT finding a significant reduction in reading time without a loss of diagnostic performance compared to DBT alone. The effect of reduced reading time was consistently shown by novice and experienced readers. The benefit of adding CAD to DBT+s2DM may have the potential to reduce reading time; however, limited research has been conducted in this area and needs further exploration. For an area of research as complex as computer aided cancer detection, many more studies will be needed before recommendations can be made regarding its future utility in population-based breast screening programs.

4.3.1. Reading time

Literature reviews

Rocha García & Fernández (2019) recognised DBT+s2DM's potential contribution to improved cancer detection but the increased reading time was cited as a major limitation. For example, a breast with tissue thickness of five centimetres, using a slice thickness of 1 mm, creates 100 images. This led to an increase in the reading time of 35 to 70 percent compared to FFDM. Despite current longer time interpretation time, the authors observed that improvements might be possible with the addition of CAD, which could reduce the need for double reading protocols in the future, however little information was provided specifically regarding this process (although we report on a study by Chae et al., 2019, which explores the role of CAD).

In Chong at al.'s (2019) narrative literature review, the authors also considered that DBT interpretation time is expected to decrease with reader's increasing experience with DBT. Additionally, with the development of new presentation modes, including automated "slabbing" of DBT sections to create thicker overlapping sections the interpretation time should be further reduced.

DBT+s2DM compared to FFDM+DBT and/or FFDM alone

RCT

Papers reporting on data from the first year of the To-Be-1 trial reported a significantly longer times required for reading and consensus. In a sub-study (interim analysis), Aase et al. (2019) reported initial reading time of 1 minute 11 sec for DBT+s2DM cases compared with 41 sec for FFDM cases (p<.01). A significant increase in time was also seen for consensus reading (3 min 12 sec with DBT+s2DM compared to 2 min 12 sec for FFDM; p<.01); however they also observed a decrease in the rate of cases discussed at consensus (6.3 percent for DBT+s2DM; 7.4 percent for FFDM; p=.03) suggesting improved interrater consensus.

Hofvind et al. (2019), reporting on the main analysis from this RCT, noted initial reading times of 1 min 6 sec with DBT+s2DM compared to 39 sec with FFDM (p<.0001). Consensus reading was 2 min 51 sec compared to 2 min 4 sec (p<.0001) and a consensus rate of 6.3 percent compared to 7.4 percent (p=·0004). Hofvind et al. also reported individual reader times of 39 sec to 2 min 42 sec for DBT+s2DM compared to 13 sec to 3 min 2 sec for FFDM. Similarly, in Moger et al.'s 2019 cost analysis study of the To-Be-1 RCT, similar differences were found (initial reading - 2 min 12 sec compared to 1 min 19 sec, p<.001; consensus was 2 min 50 sec compared to 2 min 4 sec, p<.001, and consensus rate was 6.6 percent compared to 7.6 percent, p=.001).

Aase et al. (2019) reported that time spent on screen reading and consensus was observed to be lowest in the first five to eight months after the start of the trial, but it increased during the 8-12 months period (see *Table 22*). This was thought to have been impacted by the timing of a workshop in which cancer cases that had been dismissed by one of the two readers were reviewed as a part of a quality assurance. The authors believed this may have contributed to readers deliberating longer at screen reading during the 8-12-month period. Other factors that were observed were a significant decreasing trend of consensus with reading volume increase during the trial for FFDM, but not for DBT. The volume of screen reads prior to the trial did not show any correlation with either consensus or recall rate, either for DBT or FFDM. As noted in

section 4.2, while a difference in reading time is anticipated for DBT compared to FFDM, there are potential factors that could be compounding the difference including reader training/experience with DBT and s2DM, availability of previous DBT images and potentially reader preferences.

Variation in interpretation times, consensus times, consensus rates and recall rates varied over the first 12 months of the study (see *Table 22*). Learning effect may have contributed to some of this fluctuation, with Aase citing the potential influence of a workshop during the seventh and eighth month period; however, Aase et al. also noted that months five to eight were summer months where there were fewer women being screened , resulting in lower power in the estimate.

Table 22: Interpretation based on data from Aase et al. (2019)

	Interpretation time DBT+s2DM vs FFDM	Consensus time DBT+s2DM vs FFDM	Consensus rates as % DBT+s2DM vs FFDM	Recall rates at % DBT+s2DM vs FFDM
1-4 months	DBT+s2DM: 1 min 18 sec FFDM: 42 sec *p*<.01	DBT+s2DM: 3 min 31 sec / 3 min 14 sec FFDM: 2 min 08 sec / 1min 48 sec *p*<.01	DBT+s2DM: 6.5 FFDM: 7.2 *p*=.35	DBT+s2DM 3.0 FFDM: 3.6 *p*=.25
5-8 months	DBT+s2DM: 56 sec/ 46 sec FFDM: 33 sec / 21 sec *p*<.01	DBT+s2DM: 2 min 45 sec / 2 min 14 sec FFDM: 1 min 54 sec / 1 min 42 sec *p*<.01	DBT+s2DM: 5.3 FFDM: 5.7 *p*=.67	DBT+s2DM: 2.6 FFDM: 2.0 *p*=.28
9-12 months	DBT+s2DM: 1 min 11 sec/ 54 sec FFDM: 45 sec/ 27 sec *p*<.01	DBT+s2DM: 3 min 06 sec / 2 min 39 sec FFDM: 2 min 21 sec / 2 min 05 sec *p*<.01	DBT+s2DM: 6.8 FFDM: 8.3 *p*=.03	DBT+s2DM: 3.1 FFDM: 4.4 *p*<.01

Prospective trials embedded in population screening programs

Data from the Verona pilot evaluation (Caumo et al., 2018) reported that 38.5 DBT+s2DM interpretations were completed on average per hour compared with 60 interpretations per hour for FFDM. In the Verona study, if one reader recorded a positive result, the result was considered positive as no arbitration was performed, which is of interest as the authors report a decrease in discordant referral recommendations for cancers from 28.2 percent with FFDM to 7.1 percent for DBT+s2DM (*p*=.0002). This suggests significantly higher levels of agreement between readers of DBT+s2DM compared to FFDM alone. As with the majority of other studies DBT+s2DM in this study, despite requiring increased reader time, did deliver significantly better performance in cancer detection (CDR and PPV_1) than FFDM.

The Cordoba population based prospective trial (Romero Martín et al., 2018) determined a mean interpretation time of 25s for FFDM, 61s for DBT+s2DM and 67s for DBT+s2DM+FFDM, an increase of 59 percent and 62.7 percent, respectively. The authors proposed that while interpretation time was significantly longer for DBT than FFDM, the increased time is acceptable given the significant increase in cancer detection and decrease in recalls that the study recorded. They also supported the exploration of single reading DBT+s2DM as an alternative to double reading of 2D mammography.

Results from the Maroondah pilot (Houssami et al., 2019) reported a mean reading time of 2 mins 10 sec for DBT+s2DM (IQR: 46 sec - 1 min 45 sec) compared to 43 sec for FFDM (IQR: 10-29 sec). DBT+s2DM reading was about three times as long as for FFDM and was longer than reported by other authors. The authors concluded that it was feasible to implement DBT but the disadvantages of substantially longer screen reading times need to be considered when making decisions about larger trials of DBT or screening policy.

FFDM alone compared to FFDM alone

RCT

The Reggio Emilia RCT (Pattacini et al., 2018) investigated the difference in reading and consensus time between DBT alone compared with FFDM alone. They reported times for negative results with DBT as 56 sec (IQR: 43–77 sec), compared with 34 sec for FFDM alone (IQR: 23–51 sec) (p=.01). The authors claimed that the difference in reading time had almost disappeared for positive results, but they do not provide evidence to support this claim. Consensus time (referred to as time to for recall decision) was 1 min 54s (IQR, 1min 7 s – 2min 56s) for DBT alone compared with 1 min 44 sec (IQR, 1 min 10s –2 min 30 s) (p=.54) for FFDM.

4.3.2. The potential role of CAD

In *Allen + Clarke*'s previous literature reviews, CAD was mentioned; however, the information available in these studies was limited (i.e., one study used CAD for the FFDM images as CAD was not available for DBT and the other noted its use but made no further comment). Since then, two studies have reported more information on the potential role that CAD could play in reducing interpretation time without affecting program performance.

Prospective trials embedded in population screening programs

The OTS trial reported by Skaane et al. (2019) compared FFDM+CAD to FFDM alone. The authors noted that this study provides another 'data point' regarding the utility of CAD in breast screening, but that the estimated changes in both sensitivity and specificity were small and nonsignificant and did not match the increases seen with other uses of DBT. Results were in relation to sensitivity were FFDM: 54.1 percent and FFDM+CAD: 56.2 percent. Reported changes in specificity were 94.2 percent for FFDM and 94.2 percent for FFDM+CAD.

Retrospective analysis

A study by Chae et al. (2019) reviewed a cancer enriched sample of 100 DBT cases (70 with and 30 without breast cancers) to determine if the addition of CAD to DBT could reduce the average reading time of DBT alone compared with DBT+CAD. The study found that DBT+CAD significantly reduced reading time without a loss of diagnostic performance in cancer detection:

- DBT+CAD: 1 min 2 sec (SD, 34.38 sec) compared to
- DBT alone: 1 min 12 sec (SD, 37.54 sec) (p<.001).

The average difference in reading time was a statistically significant decrease by 10.04 ± 1.85 sec (p<.001) on using CAD, providing 14 percent decrease in time. The effect of reducing the reading time was consistently shown by novice (p=.007) and experienced readers (p<.001).

Table 23: Interpretation time

Study	DBT+s2DM (time in min/sec)	FFDM+DBT (time in min/sec)	FFDM alone (time in min/sec)	CAD (time in min/sec)
DBT+s2DM compared to FFDM alone				
RCT				
Aase et al. (2019) To-Be-1 (sub-study)	Initial read: 1 min 11 sec Consensus: 3 min 12 sec Consensus rate = 6.4%	Initial read: 41 sec (p<.01). Consensus: 2 min 12 sec (p<.01), Consensus rate = 7.4%, (p=.03)	NA	NA
Hofvind et al. (2019) To-Be-1 (main analysis)	Initial read: 1min 6 sec (IQR 33–78) Individual range: 39 sec to 2 min 42 sec Consensus: 2 min 51 sec (IQR 1 min 50 sec) Consensus rate 6.3%	Initial read: 39 (IQR 13–44) (p<.0001). Individual range: 13 sec to 3 min 2 sec Consensus: 2 min 4 sec (IQR 1 min 11 sec) (p<.0001) Consensus rate 7.4%, (p=.0004)	NA	NA
Moger et al. (2019) To-Be-1 Trial (sub-study)	Initial read: 2 min 12 sec Consensus: 2 min 50 sec Consensus rate: 6.6%	Initial read: 1 min 19 sec (SD 94) (p<.001) Consensus: 2 min 4 sec (SD 2 min 4) (p<.001), Consensus rate: 7.6% (p=.001)	NA	NA
Prospective studies embedded in a population-based screening program				
Houssami et al. (2019) Maroondah	Initial read: 130 sec (median: 1 min 7 sec (IQR) 46–105)	Initial read: 43 sec (median: 16 sec (IQR, 10–29)	NA	NA
Caumo et al. (2018) Verona	38.5 interpretations per hour	60 interpretations per hour	NA	NA
Romero Martín et al. (2018) Cordoba	Initial read: 61 sec	Initial read: 25 sec	NA	NA
FFDM+DBT compared to FFDM alone				
RCT				
Pattacini et al. (2018) Reggio Emilia	NA	-ve results: 56 sec (IQR, 43–77) Consensus: 114 sec (IQR, 67–176)	-ve results: 34 sec (IQR, 23–51) (p=.01) Consensus: 104 (IQR, 70–150) (p=.54)	NA
CAD				
Retrospective analysis				
Chae et al. (2019)	NA	NA	DBT alone: 1 min 2 sec (SD 34.38 sec) (p<.001)	1min 12 sec (SD, 37.54 sec)

4.4. Cost

Implementation of DBT into a screening program is associated with both capital upgrade and operational costs including purchasing DBT-capable units and associated equipment, software licenses, storage, data transmission and any additional costs associated with increased reading time.

Key findings from *Allen + Clarke*'s 2018 literature reviews on tomosynthesis in screening and the assessment of screen-detected abnormalities

The previous *Allen + Clarke* literature reviews found that there was limited information available about the actual costs associated with either the implementation or use of DBT in a screening setting. No costs were available for the Australian setting. Some available articles described the kinds of costs that may be incurred with the use of DBT (such as capital upgrade expenditure, people resourcing costs, etc.) but did not provide financial estimates. These articles also did not cover the range of possible ways that DBT could be integrated (either as an adjunct to FFDM, with s2DM or alone. in one projection or two). That said, some papers provided modelled analysis estimations for implementation: the US models indicated some evidence that DBT in combination with FFDM resulted in overall economic favourability when balanced against the potential improved clinical outcomes; however, the US operates a different health funding model compared to the Australian system and the analysis was of limited use in the BSA setting.

Updated findings

Literature published since 31 December 2017 relating to incremental costs associated with implementing DBT is limited to a comment on drivers in a narrative literature review (Chong et al., 2019), data from the To-Be-1 RCT set in Norway (Moger et al., 2019) and a retrospective economic modelling analysis study conducted in the US (Lowry et al., 2019). Moger et al. To-Be-1 reviewed the costs associated with DBT+s2DM; Lowry et al. looked at the costs associated with FFDM+DBT. No Australian data was identified as part of this literature review. The included studies are listed below.

Literature review

One narrative literature review: Chong et al., 2019

Randomized controlled trial

One RCT: Moger et al., 2019

Observational studies

One cost-effectiveness study: Lowry et al., 2019

Summary of key findings published between 1 January 2018 and 31 December 2019

Two studies provided costing analysis (a Norwegian RCT and a US retrospective modelling analysis). Both indicated that DBT (either instead of or in addition to FFDM) required significant additional costs on a micro and macro level and in both the short- and long-term.

They estimated that increased cost would continue to be the case irrespective of more favorable scenarios such as reduction in connectivity and data storage costs, reduction in the price of DBT capable equipment, reduction in DBT reading time, increased DBT sensitivity and reduced recall rates, and reduced charges for of DBT screening procedures. No Australian cost data was reported in the first (and only) paper from the Maroondah pilot.

While there is interest in exploring the incremental costs of DBT (+/- s2DM) compared to FFDM, there continues to be a paucity of studies investigating this issue as it pertains to population-based breast screening programs. Further studies are needed to determine if higher costs associated with the investment in machines, connectivity, data storage, and longer times for screen readings of DBT might be balanced against lower recall rates, less extensive diagnostic workup and treatment due to better or earlier detection.

DBT+s2DM compared to FFDM+DBT and/or FFDM alone

Literature review

Chong et al. (2019) proposed that the main drivers of economic value in implementing DBT at the population level were:

- the reduced cost associated with a reduction in overall recall and false positive recall rates (bearing in mind that not all studies from the larger European trials reported a reduction in recall)
- a more direct or expedited mammographic evaluation at diagnostic or problem-solving imaging, which may decrease overall cost, and
- the potential to detect cancers at an earlier stage, reducing the costs associated with later stage treatments.

Randomized controlled trial

Study Design	Increased costs per woman screened of DBT compared to FFDM (Euro €; 95%CI)
Moger et al. (2019) To-Be-1 RCT	Screening costs €8.5 (8.4 to 8.6) Adding recall costs €6.2 (4.6 to 7.9) Adding treatment costs €9.8 (–56 to 74)

As part of the To-Be-1 RCT (methodology described in *section 3.1.1*), Moger et al. (2019) undertook a detailed comparison of incremental Year One cost differences between DBT+s2DM compared to FFDM alone. They considered the following costs: hourly wages of radiologists, imaging equipment, image storage and connectivity, radiology reimbursement tariff, out of pocket fees, in-hospital treatment (for example, planning meetings and 15 radiation therapy treatments), and medication (chemotherapy treatments, endocrine treatment, immune therapy and other medications). Moger et al. analysed the cost difference per screened woman using three steps:

1. Screening components only
2. Adding recall assessment costs, and
3. Adding treatment costs.

Variation in examination and reading time was observed between radiographers/radiologists so adjustment was made for radiologist fixed effects. Short-term screening performance measures were not covered in this analysis (such as the impact of reduced recall rates observed in this RCT, see the analysis from Hofvind et al.'s 2019 paper, *section 3.4.1*). CDR did not differ between the two imaging modalities. Additionally, only direct healthcare costs were included (i.e., all women who required treatment completed it as per recommended in the national guidelines, and follow-up was until the end of treatment for any detected tumours).

Moger et al. (2019) applied three scenarios to the data to determine if these significantly impacted cost differences:

- A 50 percent reduction in storage and connectivity costs
- A 30 percent reduction in additional price of tomo-equipped mammograph, and
- No additional examination time plus a 20 percent reduction in total reading time.

Results from Moger et al.'s study indicated that the incremental cost of equipment, examination and reading time increased when DBT+s2DM was used compared to FFDM, costing €8.5[31] (95%CI: 8.4–8.6) more per screened woman. The cost per woman screened was still higher after accounting for the decrease costs associated with recall assessment (bearing in mind that this RCT reported a reduction in recall): €6.2[32] (95%CI: 4.6–7.9) more per screened woman. Reductions in either examination and reading times, price of DBT equipment or price of IT storage and connectivity did not change these results.

Adding treatment costs (step 3) resulted in very wide confidence intervals, making it difficult to draw definitive conclusions. The additional costs of using DBT+s2DM were €9.8[33] (95%CI: –56, 74). An additional finding was that performing biopsy at recall, radiation therapy and chemotherapy was more frequent for women screened with DBT, which again increased costs. Overall program impacts were an incremental increase of €1.9M[34] over one year.

While this study provides useful information about the costs of implementing DBT+s2DM into a biennial double-read screening program, there continues to be a need for Australian-based cost effectiveness studies that look at Australian breast cancer prevalence/incidence rates, CDR within the BSA program, recall rates, false positive recall rates, etc. (i.e., an Australian cost-effectiveness study).

FFDM+DBT compared to FFDM alone

Cost-effectiveness study

Lowry et al. (2019) undertook a comparative modelling analysis of three established breast cancer microsimulation models to predict long term impact of integrating DBT into breast cancer screening practices in the US. The study estimated the incremental effect of DBT on breast cancer mortality, false positive screens, quality-adjusted life years (QALYs), costs, and cost-effectiveness of breast cancer screening at a population level. The three established Cancer Intervention and

[31] AUD 13.76 converted on 9 December 2019.
[32] AUD 10.03 converted on 9 December 2019.
[33] AUD 15.86 converted on 9 December 2019.
[34] AUD 3,048,359 converted on 9 December 2019

Surveillance Modelling Network (CISNET) models were Model D (Dana Farber Cancer Institute), Model G-E (Georgetown University Medical Centre and Albert Einstein College of Medicine) and Model W (University of Wisconsin and Harvard Medical School). All models used screening utilisation data from the National Health Interview Survey and the Breast Cancer Surveillance Consortium. Lowry et al. used performance data for FFDM+DBT and FFDM derived from observational data provided by the PROSPR consortium. Outcomes are reported per 1000 simulated US women aged 40-80 years in the year 2011 followed for the rest of their lives. The two screening scenarios were FFDM alone and FFDM+DBT (although the paper is unclear whether the dual acquisition mode was used in the modelling). Screening patterns, breast density distribution and adjuvant therapy and effectiveness were based on 2011 levels for the duration of the study. This may result in some slightly out of date costs (especially for adjuvant therapies which are now available, but which may not have been in 2011).

Lowry et al. compared a four percent improvement in sensitivity against no improvement in their modelling analysis, additionally they applied 50-150 percent to disutilities due to screening and diagnostic recall.

Results from the modelling analysis indicated only small improvements in health outcomes when the four percent sensitivity improvement was applied: a reduction in breast cancer deaths by 0.16 to 0.26/1000 women, an increase in life years by 2.17 to 4.36 per 1000 women, and an increase in QALYs by 3.61 to 4.97 per 1000 women. Similarly, with a four percent sensitivity improvement applied, additional costs of implementing DBT were as follows:

- USD 130,533 to 156,6248[35] per QALY per 1000 women screened (compared to USD 395,553 to 445,722[36] at zero percent improvement), and
- Incremental Cost Effectiveness Ratios (ICERs) improved if disutilities were applied for screening and diagnostic evaluation but generally remained high at base case values at:
    - 50 percent: USD 301,529 to 530,621[37]
    - 150 percent of USD 144,121 to 180,820[38].

Lowry suggested that with a predicted improvement in DBT specificity, reducing false positive by up to 26 percent, and a potential lowering of costs for individual screening examinations, that the cost effectiveness of DBT improved; however, as this analysis was undertaken in the US, which does not have a comparable health care system, medical rebate structure or screening program the findings are of interest but are of limited value to BSA.

[35] AUD 191,053 to 229,23 converted at 9 December 2019.
[36] AUD 578,946 to 652,375 converted at 9 December 2019.
[37] AUD 441328 to 776,636 converted at 9 December 2019.
[38] AUD 210,940 to 264,654 converted at 9 December 2019.

REFERENCES

Aase, H. S., Holen, Å. S., Pedersen, K., Houssami, N., Haldorsen, I. S., Sebuødegård, S., … Hofvind, S. (2019). A randomized controlled trial of digital breast tomosynthesis versus digital mammography in population-based screening in Bergen: Interim analysis of performance indicators from the To-Be-1 trial. *European Radiology*, *29*(3), 1175–1186. https://doi.org/10.1007/s00330-018-5690-x

Ambinder, E. B., Harvey, S. C., Panigrahi, B., Li, X., & Woods, R. W. (2018). Synthesized Mammography: The New Standard of Care When Screening for Breast Cancer with Digital Breast Tomosynthesis? *Academic Radiology*, *25*(8), 973–976. https://doi.org/10.1016/j.acra.2017.12.015

Bahl, M., Pinnamaneni, N., Mercaldo, S., McCarthy, A. M., & Lehman, C. D. (2019). Digital 2D versus Tomosynthesis Screening Mammography among Women Aged 65 and Older in the United States. *Radiology*, *291*(3), 582–590. https://doi.org/10.1148/radiol.2019181637

Bernardi, D., Gentilini, M. A., De Nisi, M., Pellegrini, M., Fantò, C., Valentini, M., … Houssami, N. (2019). Effect of implementing digital breast tomosynthesis (DBT) instead of mammography on population screening outcomes including interval cancer rates: Results of the Trento DBT pilot evaluation. *Breast (Edinburgh, Scotland)*. https://doi.org/10.1016/j.breast.2019.09.012

Bernardi, D., Li, T., Pellegrini, M., Macaskill, P., Valentini, M., Fantò, C., … Houssami, N. (2018). Effect of integrating digital breast tomosynthesis (3D-mammography) with acquired or synthetic 2D-mammography on radiologists' true-positive and false positive detection in a population screening trial: A descriptive study. *European Journal Of Radiology*, *106*, 26–31. https://doi.org/10.1016/j.ejrad.2018.07.008

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.

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

Bernardi D, Ciatto S, Pellegrini M, Anesi V, Burlon S, Cauli E,et al. (2012) Application of breast tomosynthesis in screening: incremental effect on mammography acquisition and reading time.Br J Radiol. 85:e1174---8.57.

Caumo, F., Zorzi, M., Brunelli, S., Romanucci, G., Rella, R., Cugola, L., … Houssami, N. (2018). Digital Breast Tomosynthesis with Synthesized Two-Dimensional Images versus Full-Field Digital Mammography for Population Screening: Outcomes from the Verona Screening Program. *Radiology*, *287*(1), 37–46. https://doi.org/10.1148/radiol.2017170745

Chae, E. Y., Kim, H. H., Jeong, J.-W., Chae, S.-H., Lee, S., & Choi, Y.-W. (2019). Decrease in interpretation time for both novice and experienced readers using a concurrent computer-aided

detection system for digital breast tomosynthesis. *European Radiology*, *29*(5), 2518–2525. https://doi.org/10.1007/s00330-018-5886-0

Choi, J. S., Han, B.-K., Ko, E. Y., Kim, G. R., Ko, E. S., & Park, K. W. (2019). Comparison of synthetic and digital mammography with digital breast tomosynthesis or alone for the detection and classification of microcalcifications. *European Radiology*, *29*(1), 319–329. https://doi.org/10.1007/s00330-018-5585-x

Chong, A., Weinstein, S. P., McDonald, E. S., & Conant, E. F. (2019). Digital Breast Tomosynthesis: Concepts and Clinical Practice. *Radiology*, *292*(1), 1–14. https://doi.org/10.1148/radiol.2019180760

Conant, E. F., Barlow, W. E., Herschorn, S. D., Weaver, D. L., Beaber, E. F., Tosteson, A. N. A., … Sprague, B. L. (2019). Association of Digital Breast Tomosynthesis vs Digital Mammography With Cancer Detection and Recall Rates by Age and Breast Density. *JAMA Oncology*, *5*(5), 635–642. https://doi.org/10.1001/jamaoncol.2018.7078

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, P. A., Wang, A., Senapati, G. M., Ip, I. K., Lacson, R., Khorasani, R., & Giess, C. S. (2019). Comparing Tumor Characteristics and Rates of Breast Cancers Detected by Screening Digital Breast Tomosynthesis and Full-Field Digital Mammography. *AJR. American Journal Of Roentgenology*, 1–6. https://doi.org/10.2214/AJR.18.21060

Fujii, M. H., Herschorn, S. D., Sowden, M., Hotaling, E. L., Vacek, P. M., Weaver, D. L., & Sprague, B. L. (2019). Detection Rates for Benign and Malignant Diagnoses on Breast Cancer Screening With Digital Breast Tomosynthesis in a Statewide Mammography Registry Study. *AJR. American Journal Of Roentgenology*, *212*(3), 706–711. https://doi.org/10.2214/AJR.18.20255

Gennaro, G., Bernardi, D., & Houssami, N. (2018). Radiation dose with digital breast tomosynthesis compared to digital mammography: Per-view analysis. *European Radiology*, *28*(2), 573–581. https://doi.org/10.1007/s00330-017-5024-4

Hodgson R, Heywang-Kobrunner SH, Harvey SC, Edwards M, Shaikh M, Arber M, Glanville J. '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.

Hofvind, S, Holen, AS, Aase, HS, Houssami, N, Sebuodegard, S, Moger, TA, Haldorsen, IS, & Akslen, L. (2019). Two-view digital breast tomosynthesis versus digital mammography in a population-based breast cancer screening programme (To-Be-1): A randomised, controlled trial. *The Lancet. Oncology*. https://doi.org/10.1016/S1470-2045(19)30161-5

Hofvind, S., Hovda, T., Holen, Å. S., Lee, C. I., Albertsen, J., Bjørndal, H., … Skaane, P. (2018). Digital Breast Tomosynthesis and Synthetic 2D Mammography versus Digital Mammography: Evaluation in a Population-based Screening Program. *Radiology*, *287*(3), 787–794. https://doi.org/10.1148/radiol.2018171361

Honig, E. L., Mullen, L. A., Amir, T., Alvin, M. D., Jones, M. K., Ambinder, E. B., … Harvey, S. C. (2019). Factors Impacting False Positive Recall in Screening Mammography. *Academic Radiology*. https://doi.org/10.1016/j.acra.2019.01.020

Houssami, N., Lockie, D., Clemson, M., Pridmore, V., Taylor, D., Marr, G., … Macaskill, P. (2019). Pilot trial of digital breast tomosynthesis (3D mammography) for population-based screening in BreastScreen Victoria. *The Medical Journal Of Australia*. https://doi.org/10.5694/mja2.50320

Houssami, N., Bernardi, D., Caumo, F., Brunelli, S., Fantò, C., Valentini, M., … Macaskill, P. (2018). Interval breast cancers in the 'screening with tomosynthesis or standard mammography' (STORM) population-based trial. *Breast (Edinburgh, Scotland)*, *38*, 150–153. https://doi.org/10.1016/j.breast.2018.01.002

Houssami N., Bernardi D., Pellegrini M., Valentini M., Fanto C., Ostillio L., … Macaskill P. (2017). Breast cancer detection using single-reading of breast tomosynthesis (3D-mammography) compared to double-reading of 2D-mammography: Evidence from a population-based trial. *Cancer Epidemiology*, *47*((Houssami, Macaskill) Sydney School of Public Health (A27), Sydney Medical School, University of Sydney, Sydney 2006, Australia), 94–99. https://doi.org/10.1016/j.canep.2017.01.008

Houssami N., Macaskill P., Bernardi D., Caumo F., Pellegrini M., Brunelli S., … Ciatto S. (2014). 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*(10), 1799–1807. https://doi.org/10.1016/j.ejca.2014.03.017

Hovda, T., Brandal, S. H. B., Sebuødegård, S., Holen, Å. S., Bjørndal, H., Skaane, P., & Hofvind, S. (2019). Screening outcome for consecutive examinations with digital breast tomosynthesis versus standard digital mammography in a population-based screening program. *European Radiology*. https://doi.org/10.1007/s00330-019-06264-y

Johnson, K., Zackrisson, S., Rosso, A., Sartor, H., Saal, L. H., Andersson, I., & Lång, K. (2019). Tumor Characteristics and Molecular Subtypes in Breast Cancer Screening with Digital Breast Tomosynthesis: The Malmö Breast Tomosynthesis Screening Trial. *Radiology*, 190132–190132. https://doi.org/10.1148/radiol.2019190132

Lai, Y.-C., Ray, K. M., Lee, A. Y., Hayward, J. H., Freimanis, R. I., Lobach, I. V., & Joe, B. N. (2018). Microcalcifications Detected at Screening Mammography: Synthetic Mammography and Digital Breast Tomosynthesis versus Digital Mammography. *Radiology*, *289*(3), 630–638. https://doi.org/10.1148/radiol.2018181180

Lång, K. (2019). The Coming of Age of Breast Tomosynthesis in Screening. *Radiology*, *291*(1), 31–33. https://doi.org/10.1148/radiol.2019190181

Li, T., Marinovich, M. L., & Houssami, N. (2018). Digital breast tomosynthesis (3D mammography) for breast cancer screening and for assessment of screen-recalled findings: Review of the evidence. *Expert Review of Anticancer Therapy*, *18*(8), 785–791. https://doi.org/10.1080/14737140.2018.1483243

Lowry, K. P., Trentham-Dietz, A., Schechter, C. B., Alagoz, O., Barlow, W. E., Burnside, E. S., … Stout, N. K. (2019). Long-term Outcomes and Cost-effectiveness of Breast Cancer Screening with Digital

Breast Tomosynthesis in the United States. *Journal Of The National Cancer Institute*. https://doi.org/10.1093/jnci/djz184

Marinovich, M. L., Hunter, K. E., Macaskill, P., & Houssami, N. (2018). Breast Cancer Screening Using Tomosynthesis or Mammography: A Meta-analysis of Cancer Detection and Recall. *Journal Of The National Cancer Institute*, *110*(9), 942–949. https://doi.org/10.1093/jnci/djy121

Miglioretti, D. L., Abraham, L., Lee, C. I., Buist, D. S. M., Herschorn, S. D., Sprague, B. L., … Kerlikowske, K. (2019). Digital Breast Tomosynthesis: Radiologist Learning Curve. *Radiology*, *291*(1), 34–42. https://doi.org/10.1148/radiol.2019182305

Moger, TA, Swanson, JO, Holen, AS, Hanestad, B., & Hofvind, S. (2019). Cost differences between digital tomosynthesis and standard digital mammography in a breast cancer screening programme: Results from the To-Be-1 trial in Norway. *European Journal of Health Economics*. https://doi.org/10.1007/s10198-019-01094-7

Østerås, B. H., Skaane, P., Gullien, R., & Martinsen, A. C. T. (2018). Average glandular dose in paired digital mammography and digital breast tomosynthesis acquisitions in a population based screening program: Effects of measuring breast density, air kerma and beam quality. *Physics In Medicine And Biology*, *63*(3), 035006–035006. https://doi.org/10.1088/1361-6560/aaa614

Pattacini, P., Nitrosi, A., Giorgi Rossi, P., Iotti, V., Ginocchi, V., Ravaioli, S., … Campari, C. (2018). Digital Mammography versus Digital Mammography Plus Tomosynthesis for Breast Cancer Screening: The Reggio Emilia Tomosynthesis Randomized Trial. *Radiology*, *288*(2), 375–385. https://doi.org/10.1148/radiol.2018172119

Phi, X.-A., Tagliafico, A., Houssami, N., Greuter, M. J. W., & de Bock, G. H. (2018). Digital breast tomosynthesis for breast cancer screening and diagnosis in women with dense breasts—A systematic review and meta-analysis. *BMC Cancer*, *18*(1), 380–380. https://doi.org/10.1186/s12885-018-4263-3

Rocha García, A. M., & Mera Fernández, D. (2019). Breast tomosynthesis: State of the art. *Radiologia*, *61*(4), 274–285. https://doi.org/10.1016/j.rx.2019.01.002

Romero Martín, S., Raya Povedano, J. L., Cara García, M., Santos Romero, A. L., Pedrosa Garriguet, M., & Álvarez Benito, M. (2018). Prospective study aiming to compare 2D mammography and tomosynthesis + synthesized mammography in terms of cancer detection and recall. From double reading of 2D mammography to single reading of tomosynthesis. *European Radiology*, *28*(6), 2484–2491. https://doi.org/10.1007/s00330-017-5219-8

Rose, S. L., & Shisler, J. L. (2018). Tomosynthesis Impact on Breast Cancer Screening in Patients Younger Than 50 Years Old. *AJR. American Journal of Roentgenology*, *210*(6), 1401–1404. https://doi.org/10.2214/AJR.17.18839

Simon, K., Dodelzon, K., Drotman, M., Levy, A., Arleo, E. K., Askin, G., & Katzen, J. (2019). Accuracy of Synthetic 2D Mammography Compared With Conventional 2D Digital Mammography Obtained With 3D Tomosynthesis. *AJR. American Journal Of Roentgenology*, 1–6. https://doi.org/10.2214/AJR.18.20520

Skaane, P., Bandos, A. I., Niklason, L. T., Sebuødegård, S., Østerås, B. H., Gullien, R., … Hofvind, S. (2019). Digital Mammography versus Digital Mammography Plus Tomosynthesis in Breast

Cancer Screening: The Oslo Tomosynthesis Screening Trial. *Radiology*, *291*(1), 23–30. https://doi.org/10.1148/radiol.2019182394

Skaane, P., Sebuødegård, S., Bandos, A. I., Gur, D., Østerås, B. H., Gullien, R., & Hofvind, S. (2018). Performance of breast cancer screening using digital breast tomosynthesis: Results from the prospective population-based Oslo Tomosynthesis Screening Trial. *Breast Cancer Research And Treatment*, *169*(3), 489–496. https://doi.org/10.1007/s10549-018-4705-2

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

Upadhyay, N., Soneji, N., Stewart, V., & Ralleigh, G. (2018). The effect of the addition of tomosynthesis to digital mammography on reader recall rate and reader confidence in the UK prevalent screening round. *Clinical Radiology*, *73*(8), 744–749. https://doi.org/10.1016/j.crad.2018.03.013

Wahab, R. A., Lee, S.-J., Zhang, B., Sobel, L., & Mahoney, M. C. (2018). A comparison of full-field digital mammograms versus 2D synthesized mammograms for detection of microcalcifications on screening. *European Journal Of Radiology*, *107*, 14–19. https://doi.org/10.1016/j.ejrad.2018.08.004

Wasan, R. K., Morel, J. C., Iqbal, A., Michell, M. J., Rahim, R. R., Peacock, C., … Satchithananda, K. (2019). Can digital breast tomosynthesis accurately predict whether circumscribed masses are benign or malignant in a screening population? *Clinical Radiology*, *74*(4), 327.e1-327.e5. https://doi.org/10.1016/j.crad.2018.12.020

Zackrisson, S. (2019). Tomosynthesis in breast screening: Great expectations? *The Lancet. Oncology*, *20*(6), 745–746. https://doi.org/10.1016/S1470-2045(19)30287-6

Zackrisson, S., Lång, K., Rosso, A., Johnson, K., Dustler, M., Förnvik, D., … Andersson, I. (2018). One-view breast tomosynthesis versus two-view mammography in the Malmö Breast Tomosynthesis Screening Trial (MBTST): A prospective, population-based, diagnostic accuracy study. *The Lancet. Oncology*, *19*(11), 1493–1503. https://doi.org/10.1016/S1470-2045(18)30521-7

ANNEX A: 2018 DASHBOARD

The current evidence base comparing tomosynthesis to digital mammography reports that …			
		FFDM+DBT	**DBT+s2DM**
Fast	Image acquisition time	Between 3 and 25 sec to capture DBT image, which is time on top of FFDM image acquisition	Between 3 and 25 sec 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
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

ANNEX B: STUDY POPULATIONS

Screening-only	Mixed (screening + diagnostic)
Aase et al. (2019) Bahl et al. (2019) Bernardi et al. (2019) Bernardi et al. (2018) Bernardi et al. (2016) Bernardi + Houssami (2017) Bernardi et al. (2012) Caumo et al. (2018) Conant et al. (2019) Dang et al. (2019) Hofvind et al. (2019) Hofvind et al. (2018) Honig et al. (2019) Houssami et al. (2019) Houssami et al. (2018) Houssami et al. (2017) Houssami et al. (2014) Hovda et al. (2019) Fuiji et al. (2019) Gennaro et al. (2018) Johnson et al. (2019) Lai et al. (2018) Lowry et al. (2019) Marinovich et al. (2018) Miglioretti et al. (2019) Moger et al. (2019) Østerås et al. (2018) Pattacini et al. (2018) Romero Martín et al. (2018) Rose + Shisler (2018) Simon et al. (2019) Skaane et al. (2019) Skaane et al. (2018) Upadhyay et al. (2018) Wahab et al. (2018) Wasan et al. (2019) Zackrisson et al. (2018)	*NB most of these studies describe cancer characteristics detected with DBT ± s2DM or FFDM/FFDM alone or reading time* Chae et al. (2019) Choi et al. (2019) Phi et al. (2018)

ANNEX C: QUALITY ASSESSMENT FOR EACH INCLUDED SYSTEMATIC REVIEW

Marinovich et al. (2019)

	AMSTAR2 TOOL QUESTION	Answer	Comment
1	Did the research questions and inclusion criteria for the review include the components of the PICO(T/S)?	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	Medline, Embase, PreMedline, HTA, NHSSEED, SCP Journal Club, Cochrane
5	Was there duplicate study selection and data extraction?	Yes	
6	Did the review authors provide a list of excluded studies and justify the exclusion?	No	But supplementary material available online described clear inclusion and exclusion criteria
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 was used
9	Did the review authors report on the sources of funding for the studies included in the review?	No	

	AMSTAR2 TOOL QUESTION	Answer	Comment
10	If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?	Yes	Logistic regression models with random effects were used, DerSimoneon and Laird methods, PROC GENMOD for paired study data and forest plots
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	See above
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?	No	Most included studies were large.
15	Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	No	

Phi et al. (2018)

	AMSTAR2 TOOL QUESTION	Answer	Comment
1	Did the research questions and inclusion criteria for the review include the components of the PICO(T/S)?	Yes	Included studies set in both diagnostic and screening settings and which reported on at least one of four outcomes: CDR, recall rate, sensitivity, ad/or specificity); studies included at least 100 women with dense breasts
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, discordance between reviewers and consensus reached or mediated by a third reviewer

	AMSTAR2 TOOL QUESTION	Answer	Comment
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 PubMed and Scopus (Jan 2017 – May 2017) plus a manual bibliography check of included articles
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: did not contain original data, simulation studies
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	Modified QUADAS-2 used by two reviewers independently: domains used were patient selection, index test, reference standard, flow and timing, applicability
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	Random effects model (RevMan 5.3); analysis completed separately on screening and diagnostic populations; sub-group analysis completed to examine effect of covariates, modality, reading protocol and outcome; heterogeneity was quantified with I^2 for CDR and recall rate
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?	Not sure	RoB is not discussed
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

	AMSTAR2 TOOL QUESTION	Answer	Comment
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	