The Breast 27 (2016) 52–61

Contents lists available at ScienceDirect

The Breast

journal homepage: www.elsevier.com/brst

Review

Systematic review of 3D mammography for breast cancer screening

Robert Hodgson [a], Sylvia H. Heywang-Köbrunner [b], Susan C. Harvey [c], Mary Edwards [a], Javed Shaikh [a], Mick Arber [a], Julie Glanville [a, *]

[a] York Health Economics Consortium, University of York, York, UK
[b] Referenzzentrum Mammographie Munchen, Munich, Germany
[c] Johns Hopkins Medical Institute, Baltimore, USA

ARTICLE INFO

Article history:
Received 15 October 2015
Received in revised form
23 December 2015
Accepted 6 January 2016
Available online 25 March 2016

Keywords:
Tomosynthesis
Full-field digital mammography
Breast cancer screening
Tumour imaging
Mammography
Systematic review

ABSTRACT

This review investigated the relative performance of digital breast tomosynthesis (DBT) (alone or with full field digital mammography (FFDM) or synthetic digital mammography) compared with FFDM alone for detecting breast cancer lesions in asymptomatic women. A systematic review was carried out according to systematic reviewing principles provided in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. A protocol was developed *a priori*. The review was registered with PROSPERO (number CRD42014013949). Searches were undertaken in October 2014. Following selection, five studies were eligible. Higher cancer detection rates were observed when comparing DBT + FFDM with FFDM in two European studies: the summary difference per 1000 screens was 2.43 (95% CI: 1.8 to 3.1). Both European studies found lower false positive rates for individual readers. One found a lower recall rate based on conditional recall. The second study was not designed to compare post-arbitration recall rates between FFDM and DBT + FFDM. One European study presented data on interval cancer rates; sensitivity and specificity for DBT + FFDM were both higher compared to FFDM. One large multicentre US study showed a higher cancer detection rate for DBT + FFDM, while two smaller US studies did not find statistically significant differences. Reductions in recall and false positive rates were observed in the US studies in favour of DBT + FFDM. In comparison to FFDM, DBT, as an adjunct to FFDM, has a higher cancer detection rate, increasing the effectiveness of breast cancer screening. Additional benefits of DBT may also include reduced recalls and, consequently, reduced costs and distress caused to women who would have been recalled.

Contents

* Corresponding author. York Health Economics Consortium (YHEC), Enterprise House, Innovation Way, University of York, Heslington, York, YO10 5NQ, UK.
E-mail address: julie.glanville@york.ac.uk (J. Glanville).

http://dx.doi.org/10.1016/j.breast.2016.01.002

Introduction

Breast cancer is a significant cause of mortality and morbidity for women worldwide and is the most common cancer diagnosed in women, with an estimated 1.67 million new cases diagnosed worldwide in 2012 [1]. The incidence of breast cancer is highest in developed countries with an age-adjusted incidence rate of 80 per 100,000 in the European Union and 92 per 100,000 in North America [2], and it is the second most common cause of cancer death in women in developed countries [2,3].

Screening with mammography can assist in detecting breast cancer at earlier stages, which is associated with reductions in mortality [4,5]. A recent systematic review of screening programme studies found that the screening reduced mortality from breast cancer for women invited to screening by approximately 23% and for regular participants by approximately 40% [6].

Over the last decade, the majority of screening programmes have changed from two dimensional (2D) analogue mammography to full field digital mammography (FFDM). Digital mammography (DM) is associated with small increases in detection rates and reductions in the number of false positives and is therefore likely to increase the effectiveness of screening programmes [7,8]. It represents the current standard for most mammography programmes and is the comparator in this review.

Description of the intervention

Digital breast tomosynthesis (DBT) (or three-dimensional (3D) mammography) is a development of FFDM providing analysis of 3D mammographic data through a series of tomographic image slices through the breast allowing reconstruction in thin slices. This provides greater detail and addresses the challenges of overlapping tissue, which both obscures and mimics cancer. Both DBT and mammography can be performed in one or two views.

DBT can also be used as an adjunct to FFDM: this requires a second radiation exposure, increasing the dosage required for FFDM. A DBT dataset can also be used to generate so-called synthetic 2D images, avoiding the need for additional radiation exposure. When DBT generates synthetic 2D images, the total patient radiation exposure is similar to or slightly higher than FFDM [9]. The aim of this systematic review was to examine the performance of DBT for breast cancer-screening.

Rationale

Published systematic reviews have assessed the use of DBT for the detection and diagnosis of breast cancer. Lei et al. [10] examined the relative performance of DBT + FFDM compared to FFDM in women with mammographically evident breast lesions. They concluded that one view DBT + FFDM has higher sensitivity and specificity than FFDM. This meta-analysis was limited by search criteria which resulted in the inclusion of relatively small studies (the largest study included 738 patients) with a high degree of variation in design. Houssami et al. [11,12] considered the relative performance of DBT + FFDM and FFDM in detecting breast cancer. Although studies in the Houssami review demonstrated increases in cancer detection rate using DBT + FFDM over FFDM, its conclusions regarding screening were qualified due to the test setting, small numbers and inclusion of diagnostic cases with a high prevalence of cancers. Additionally, the final search date for this review was October 2012, prior to the publication of several major tomosynthesis screening trials. A 2015 editorial by Houssami [12] summarizes several recent screening studies, but this review was not performed systematically and may not be comprehensive. Further studies and reviews are required.

While there is evidence suggesting that DBT shows superior performance diagnosing breast cancer, the relative performance of DBT and FFDM for detecting cancer in an asymptomatic, screening population has not been fully explored and researchers have suggested that this should be assessed by a variety of methods including a current review [11].

Review question

What is the performance of DBT (alone or in combination with FFDM or synthetic DM) for detecting breast cancer compared with FFDM alone when screening asymptomatic women?

Methods

This systematic review was carried out according to the systematic review guidance provided in the Cochrane Handbooks [13,14]. A protocol was developed *a priori* and was registered with PROSPERO (CRD42014013949) on 29 September 2014. The searches were performed and concluded in October 2014.

Eligibility criteria

Prospective studies or retrospective studies with 1000+ participants, evaluating the following comparisons were eligible for the systematic review:

- FFDM alone compared to DBT alone;
- FFDM alone compared to DBT + FFDM;
- FFDM alone compared to DBT + DBT-generated 2D images.

Studies were required to have been performed on systems possessing a CE mark or US Food and Drug Administration (FDA) approval.

Studies were eligible for inclusion in the review if they evaluated women participating in a breast cancer screening programme or who were undergoing opportunistic mammography screening.

Studies evaluating women meeting any of the following criteria were excluded:

- Having a previous diagnosis of and treatment for breast cancer;
- Presenting with symptoms of breast cancer or having been referred for examination (because of the detection of a possible lump);
- Having been recalled for diagnosis or further testing following a screening mammogram.

Studies were also ineligible if they met any of the following criteria:

- Compared DBT with (2D) analogue/film mammography;
- Evaluated DBT systems for the purpose of technological development;
- Reported results in languages other than English;
- Were reported only as conference abstracts;
- Were conducted before 2008 (before 2008 no system had a CE mark or FDA approval).

Reference standard

The reference standard for the positive cases of cancer was histological results confirmed by biopsy or surgical resection.

The reference standard for the negative cases was any follow-up period, where reported. The follow-up period is important to determine whether any recall rate reduction leads to an increase in missed cancers over time. However, if follow-up was not reported, studies were still eligible, although such studies did not provide information regarding absolute sensitivity.

Search strategy, selection and data extraction

Sensitive searches were conducted in relevant international databases of published research (full details are provided in the Supplementary Appendix) up to October 2014. Reference lists of relevant papers retrieved by the searches were scanned for potentially eligible studies. Systematic reviews identified by the searches were checked for additional reported research not retrieved by the database searches. Citation searches were carried out on identified records.

Before proceeding to formal record selection, irrelevant records (animal studies, conference abstracts and editorials) were removed by an experienced information specialist. Two reviewers (JG, JS) independently selected records using information in the title and abstract. Records which were of unclear relevance were retained. The full documents of all potentially relevant studies were obtained and were assessed for relevance by one reviewer and checked by a second reviewer. Disagreements on relevance were resolved through discussion or consulting a third reviewer. Studies considered ineligible, based on an assessment of the full document, are listed in the Supplementary Appendix.

As recommended by the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy [13], the risk of bias of studies was assessed by two reviewers independently using 11 of the 14 mandatory items in the QUADAS-2 tool [15,16].

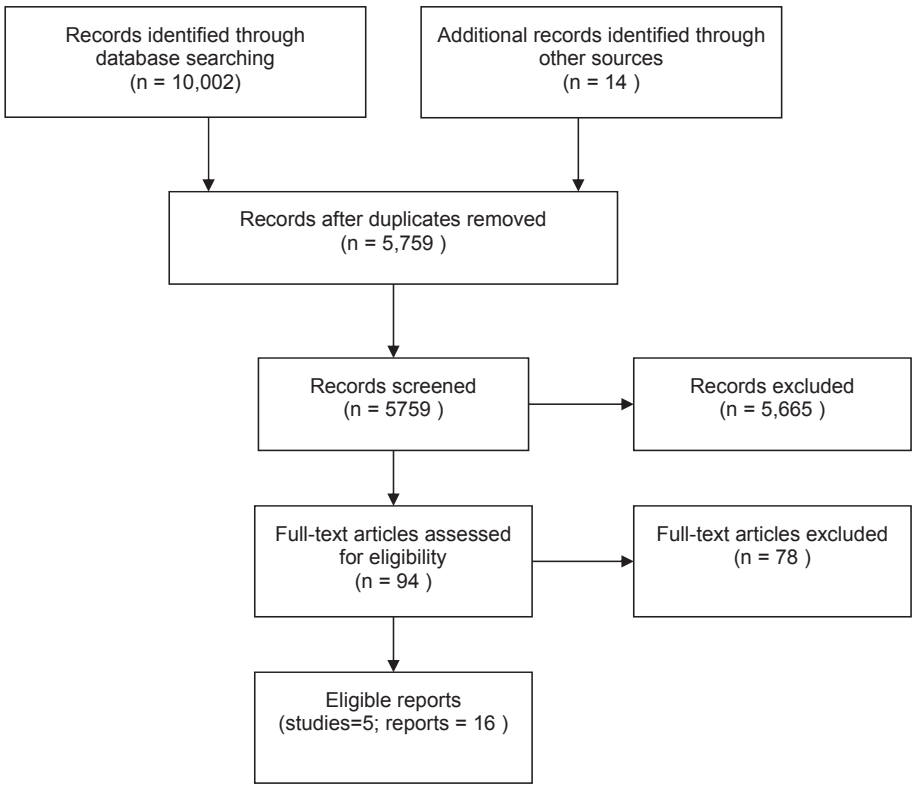

Fig. 1. Study selection process.

Table 1
Summary of study characteristics.

Study name	Country	Study design	Index test and manufacturer	Comparator test and manufacturer	Age	Inclusion criteria/ Exclusion criteria	Reference standard	Double or single reader
STORM (Ciatto 2013)	Italy	Prospective, fully paired design	Integrated 2D and 3D mammography; Hologic − Selenia Dimensions	2D mammography; Hologic	Median age: 58 years (IQR 54–63, range 48–71)	Women aged > 48 years attending a population-based breast cancer screening programme. Participants were asymptomatic women at standard (population) risk for breast cancer.	Excision histology for those patients who received surgery and complete outcome assessment with and without needle biopsy in those who did not receive surgery.	Double reader
STORM (Houssami 2014)	Italy	Prospective, fully paired design	Integrated 2D and 3D mammography; Hologic − Selenia Dimensions	2D mammography; Hologic	Median age: 58 years Women aged ≥48 years	Women aged > 48 years attending a population-based breast cancer screening programme. Participants were asymptomatic women at standard (population) risk for breast cancer.	Excision histology for those patients who received surgery and complete outcome assessment with and without needle biopsy in those who did not receive surgery.	Single and retrospective double reading algorithm
Destounis 2014	US	Retrospective review	Integrated 2D and 3D mammography; Hologic − Selenia Dimensions	2D mammography; Three manufacturers: 1. Selenia/Dimensions, Hologic; 2. Senographe Essential, GE; 3. Fuji CRm, FUJIFILM	1. Average age FFDM only group: 59 years (range: 30–90 years) 2. Average age DBT + FFDM group: 59 years (range: 36–92 years)	Women aged > 30 years attending a New York screening mammography centre.	Biopsy	Double reader
Friedewald 2014	US	Retrospective review	Integrated 2D and 3D mammography; Hologic − Selenia Dimensions	2D mammography; Hologic	1. Average age for digital mammography alone: 57 years (range of means from 13 sites, 54.4–60.5 years) 2. Average age for digital mammography + tomosynthesis: 56.2 years (range, 52.6–59.7 years)	Women were enrolled from 13 different radiology sites in the US.	Biopsy	Single reader
Lourenco 2014	US	Retrospective review	2D mammography; GE Medical and 3D mammography Hologic − Selenia Dimensions	2D mammography; GE Medical	1. Average age for DM: 54.6 years ± 10.7 (range, 29.4–90.6 years) 2. Average age for DBT: 55.3 years ± 10.8 (range, 30.9 −89.4 years)	NR	Biopsy	Single reader
OTST (Skaane 2013A)	Norway	Prospective, fully paired design	Integrated 2D and 3D mammography; Hologic − Selenia Dimensions	Integrated	Age range: 50–69 years	Women (aged 50 −69 years) who participated in the biennial Oslo breast cancer screening programme. Potential candidates were selected on the basis of the availability of appropriate staff.	Biopsy	Single reader

(*continued on next page*)

Table 1 (*continued*)

Study name	Country	Study design	Index test and manufacturer	Comparator test and manufacturer	Age	Inclusion criteria/ Exclusion criteria	Reference standard	Double or single reader
OTST (Skaane 2013B)	Norway	Prospective, fully paired design	Integrated 2D and 3D mammography; Hologic – Selenia Dimensions	2D mammography and 2D mammography plus CAD; Hologic - Dimensions	Average age: 59.3 years (range: 50–69 years)	Women (aged 50 –69 years) who participated in the biennial Oslo breast cancer screening programme. Potential candidates were selected on the basis of the availability of appropriate staff.	Biopsy	Double reader
OTST (Skaane 2014)	Norway	Prospective, fully paired design	Integrated 2D and 3D mammography; Hologic – Selenia Dimensions	Synthesized 2D and 3D mammography; Hologic – Selenia Dimensions	Average age; 59.2 years (range 50–69 years)	Women (aged 50 –69 years) who participated in the biennial Oslo breast cancer screening programme. Potential candidates were selected on the basis of the availability of appropriate staff.	Biopsy	Double reader

CAD – computer-aided detection; NR – not reported.

Details of eligible studies were extracted and summarised using an Excel data extraction template. Data were extracted by one reviewer and checked by a second reviewer.

Discrepancies between reviewers were resolved through discussion or by consulting a third reviewer.

Data analysis

Analysis was conducted using Stata 12.0 and performed in compliance with the methods and techniques described in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy [13]. Fixed effect analysis was carried out where statistical heterogeneity was low and the random effects model was used where moderate heterogeneity was observed. Where high heterogeneity was observed, a combined summary estimate was not performed and a narrative exploration of differences was conducted.

It was pre-planned to investigate publication bias using funnel plots if ten or more studies were identified. However, the small number of studies identified meant that this was not possible.

Results

Search results

Study selection is presented in Fig. 1. 10,016 records were identified from the searches and after removing duplicates, 5759 records were assessed for relevance. The full documents of 94 potentially eligible studies were obtained and five studies (reported in 16 documents) met the review eligibility criteria. The documents excluded based on the full text are listed in the Supplementary Appendix.

Characteristics of the included studies

Table 1 summarises the characteristics of the eligible studies. Two studies were conducted in Europe; the Oslo Tomosynthesis Screening Trial (OTST) [9,17,18] was conducted in Oslo, Norway and the STORM study [19–22] was carried out in Italy. The other three studies (Destounis 2014 [23], Lourenco 2014 [24] and Friedewald 2014 [25–31]) were conducted in the US. The largest study, Friedewald 2014 [26], was a multicentre study that enrolled women at 13 centres.

OTST [9,17,18] and STORM [19–22], were undertaken within population-based biannual mammography screening programmes. They used a prospective fully paired design in which screened women were invited consecutively to participate and undergo two-view FFDM and two-view DBT. Both used an independent double reader process, although the processes used to decide which women to recall were different. In the OTST study [9,17,18] consensus double reading was undertaken of slightly different data sets. The study had four reading arms and to generate double reading data the readings from the FFDM alone reading arm was combined with the FFDM and computer aided detection (CAD) reading arm, and for the combination mode, the results of the FFDM and DBT arm were combined with the results of the synthetic 2D mammography and DBT reading arm. The blinded interpretations were entered into the screening database. Cases with a positive interpretation by either reader were presented to a single consensus group of readers who had access to all the imaging information (FFDM and DBT) to inform their decision about whether to recall the patient. In the STORM study [19–22], women were recalled if either reader recorded a positive screen, so no process of arbitration was used in the decision to recall. The comparison of recall and false positive rates reported is therefore a retrospective

Table 2
Summary of QUADAS-2 risk of bias assessment.

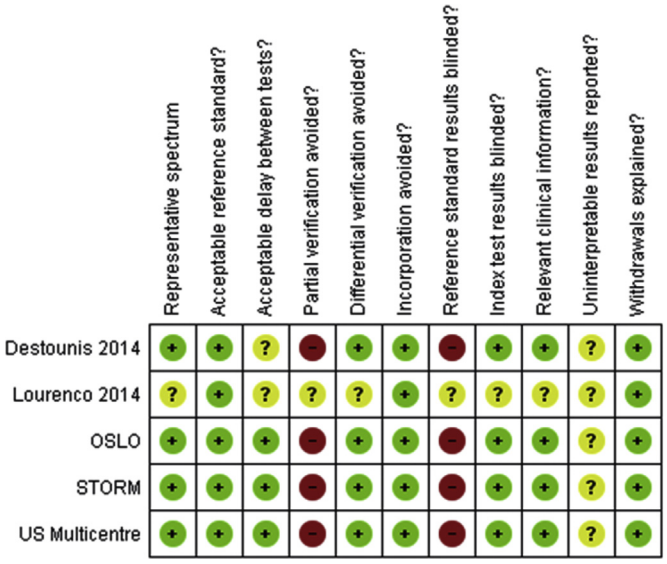

Table 3
DBT + FFDM vs. FFDM: Sensitivity and Specificity based on 1 year follow-up of STORM.

DBT + FFDM				FFDM			
Cancer detected at screening	Missed cancers + interval cancers	Sensitivity	Specificity	Cancer detected at screening	Missed cancers + interval cancers	Sensitivity	Specificity
59	6	90.77% CI (80.7%– to 96.51%)	96.49% CI (96.04% to 96.90%)	39	26	60.00% CI: (47.10% to 71.96%)	95.55% CI (95.04% to 96.01%)

conditional rate, rather than (as is typical in Europe) a double reading strategy using arbitration.

The OTST study [9,17,18] had four reading arms. Arm 1 used FFDM alone, arm 2 used FFDM plus computer aided diagnosis, arm 3 used DBT + FFDM, and arm 4 used DM synthetically-generated by DBT. The results of the study were presented in three manuscripts. One report [18] compared a single read of FFDM with a single reading of DBT + FFDM and a second [17] reported the same comparison using double reading. A third report [9] presented results of DBT + FFDM versus DBT + synthetically-generated DM.

STORM was reported in multiple manuscripts [19–22]. Ciatto et al. [21] presented results for cancer detection rates and false

Table 4
DBT + FFDM versus FFDM: false positives, recall rate, cancer detection rate, invasive cancer detection rates.

Study	DBT + FFDM				FFDM			
	False positives	Recall rate	Cancer detection rate	Invasive cancer detection rate	False positives	Recall rate	Cancer detection rate	Invasive cancer detection rate
European studies								
STORM	254/7294[a] (3.5%)	313/7294[a] (4.3%)	59/7294 (0.81%)	52/7294 (0.71%)	322/7294 (4.4%)	362/7294 (5.0%)	39/7294 (0.53%)	35/7294 (0.48%)
OTST *single reading*	670/12,621[b] (5.31%)	351/12,621[b] (2.78%)	101/12,621 (0.80%)	81/12,621 (0.64%)	771/12,621[b] (6.11%)	265/12,621[b] (2.1%)	77/12,621 (0.61%)	56/12,621 (0.44%)
OTST *double reading*	1057/12,621[b] (8.5%)	463/12,621[b] (3.67%)	119/12,621 (0.94%)	94/12,621 (0.74%)	1286/12,621[b] (10.3%)	365/12,621[b] (2.9%)	90/12,621 (0.71%)	67/12,621 (0.53%)
US studies								
Destounis 2014	19/524 (3.63%)	22/524 (4.20%)	3/524 (0.57%)	1/524 (0.19%)	58/524 (11.07%)	60/524 (11.45%)	2/524 (0.38%)	1/524 (0.19%)
Lourenco 2014	767/12,921 (5.94%)	827/12,921 (6.40%)	60/12,921 (0.46%)	30/12,921 (0.23%)	1107/12,577 (8.80%)	1175/12,577 (9.3%)	68/12,577 (0.54%)	41/12,577 (0.33%)
Friedewald 2014	14,591/173,663 (8.40%)	15,541/173,663 (8.95%)	950/173,663 (0.55%)	707/173,663 (0.41%)	28,519/281,187 (10.14%)	29,726/281,187 (10.57%)	1207/281,187 (0.43%)	815/281,187 (0.29%)

[a] False positives and recalls for the DBT + FFDM arm of the STORM trial were calculated using positive integrated DBT and FFDM as a condition to recall (i.e. exams which were positive based on FFDM, but not DBT, would not be recalled).

[b] False positives for the OTST were calculated as the number of participants without a verified cancer who were referred to arbitration. Recalls were determined based on cases sent for further evaluation after arbitration, during which FFDM and DBT information was available for all cases (including those sent to arbitration based on FFDM data alone).

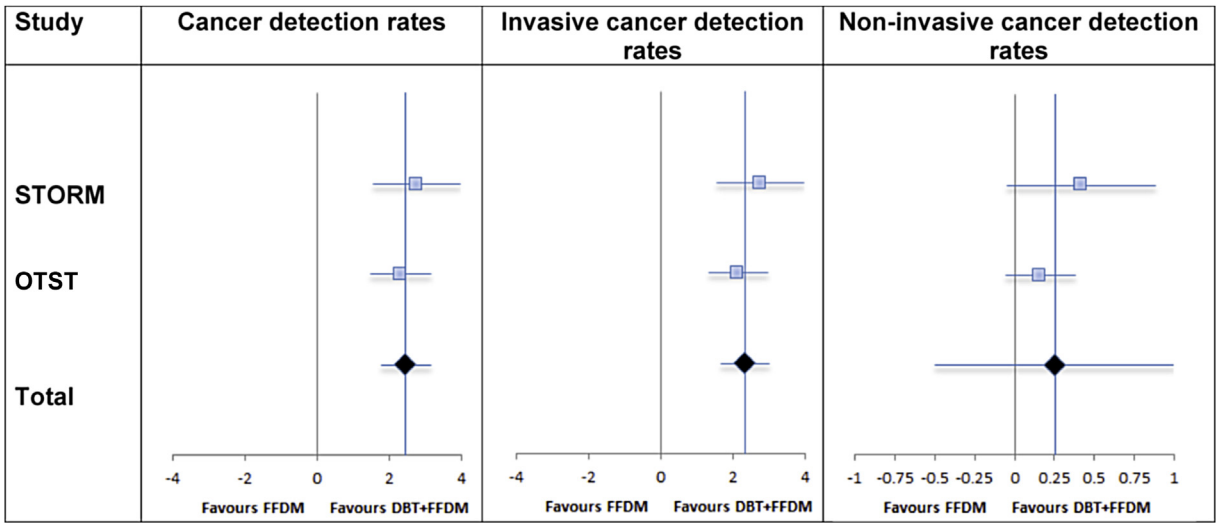

Fig. 2. DBT + FFDM vs. FFDM: cancer and invasive cancer detection rates per 1000 screens (European studies).

positive rate and Houssami et al. [22] reported results for interval cancer.

All of the US studies [23–31] were retrospective reviews in which screened women received either FFDM, or DBT + FFDM. The US multicentre study was also reported in a number of additional manuscripts. Most notable is Rose et al. [31] which reported a paired analysis (DBT + FFDM vs FFDM alone) of a subset of the participants. In two of the US studies (Lourenco 2014 and Friedewald 2014 [25–31]) single reader design was used, with only one radiologist reviewing the images (standard practice in the US). In the third US study (Destounis 2014 [23]) double reading of mammography images was performed.

Risk of bias in included studies

Four of the five eligible studies were rated to have low risk of bias and one (Lourenco 2014 [24]) was rated to have unclear risk of bias (Table 2). Studies were rated high risk of bias based on the domain of partial verification, because follow-up was applied only to those patients recalled. Studies were also rated high risk of bias based on blinding of the application of the reference standard. In all cases this was due to the nature of the study design, as it is impossible to blind readers to the type of mammography they are reviewing. This was not considered to represent a significant risk of bias for the outcomes presented in the studies. Lourenco 2014 [24] was rated to have unclear risk of bias on 8/11 of the criteria. All three of the US studies used a retrospective observational design, and this creates the potential for confounding bias, where observed differences between screening methods are attributable to differences between groups. However, all of the studies were conducted at the same sites, and patient groups were well balanced in terms of demographic factors.

Study results

The results of US and European studies were treated separately, because of differences in breast cancer rates, demographics and screening practices.

Results of the European studies

As there was only limited follow-up within the studies and only one study (STORM [19–22]) presented interval cancers, it was not possible to assess programme sensitivity and specificity. Table 3 presents estimated sensitivity and specificity based on the limited follow up data reported in STORM [19–22]. DBT + FFDM has higher sensitivity than FFDM: 90.77% (95% CI: 80.70% to 96.51%) compared with 60.00% (95% CI: 47.10% to 71.96%). DBT + FFDM also had higher specificity than FFDM: 96.49% (95% CI: 96.04% to 96.90%) compared with 95.55% (95% CI: 95.04% to 96.01%).

Table 4 presents cancer detection, false positive and recall rates of the two European studies comparing DBT + FFDM with FFDM. In both studies a higher cancer detection rate was observed using DBT + FFDM compared to FFDM.

A fixed effect meta-analysis of the studies (Fig. 2) gives a highly statistically significant ($p < 0.001$) summary difference in the cancer detection rate per 1000 screens of 2.43 (95% CI: 1.76 to 3.1) for all cancers. Similarly, both studies observed a higher invasive cancer detection rate, but did not report a statistically significant difference in non-invasive (in situ) cancer detection.

A fixed effect meta-analysis of the studies (Fig. 2), also gives a highly statistically significant ($p < 0.001$) summary difference in the invasive cancer detection rate per 1000 screens of 2.33 (95% CI: 1.67 to 3.00). Results based on the single reader mode are nearly identical to those using a double reader mode with fewer false positives and recalls, and higher cancer detection using DBT + FFDM.

The two European studies observed quite different results with respect to the false positive rate and the recall rate. In STORM [19–22] lower false positive and recall rates were observed when using DBT + FFDM compared to FFDM. The difference per 1000 screens for false positives was −9.3 (95% CI: −11.8 to −7.2) and for recall rate was −6.6 (95% CI: −8.7 to −4.9). In OTST [9,17,18] lower false positive rates using FFDM + DBT were found pre-arbitration, but higher false positive and recall rates were found post-arbitration. The difference per 1000 screens for false positives in pre-arbitration was −8 for FFDM + DBT versus FFDM alone. After consensus by arbitration, the difference for FFDM + DBT versus FDFM was +5.4 (95% CI: 4.2 to 6.8) for false positives per 1000 screens and +6.2 (95% CI: 4.9 to 7.7) for recalls per 1000 screens. Attempts to combine the results of these studies using meta-analysis resulted in significant heterogeneity: $I^2 = 99\%$ for the false positives and $I^2 = 89\%$ for the recall rate. A summary effect was, therefore, not calculated.

Results of the US studies

Table 4 presents the results of the US studies. The large multicentre study, Friedewald 2014 [25–31] found a highly statistically significant difference per 1000 screens in favour of DBT + FFDM over FFDM: 1.21 (95% CI: 0.82 to 1.63). The small Destounis 2014 study [23] found a 1.91 difference (95% CI: −6.43 to 10.25; NS) in favour of DBT + FFDM and the small Lourenco 2014 study [24] found a cancer detection rate difference of −0.76 (95% CI: −2.5 to 0.97; NS) favouring FFDM.

A statistically significantly higher invasive cancer detection rate in favour of DBT + FFDM was observed in the large Friedewald 2014 study [25–31]: the difference per 1000 screens was 1.20 (95% CI: 0.80 to 1.60). In the smaller studies, a lower rate (−0.94 (95% CI: −2.2 to 0.35)) in favour of DBT + FFDM was found in Lourenco 2014 [24] and no difference in number of invasive cancers was found in Destounis [23]. These differences mean that attempts to combine the results using meta-analysis resulted in significant heterogeneity (I^2 = 56%) for the cancer detection rate and for the invasive cancer detection rate (I^2 = 79%). A summary effect was therefore not calculated. None of the US studies reported a statistically significant difference in non-invasive cancer detection rates between DBT + FFDM and FFDM alone.

In all of the US studies the proportion of false positives observed was higher in the FFDM group. However, the magnitude of the difference in false positives rates varied across the studies, so they were not combined using meta-analysis. In the large US multicentre study (Friedewald 2014 [25–31]), a modest reduction in the number of false positives was observed: a difference per 1000 screens of −17.4 (95% CI: −15.6 to −19.2) in favour of DBT + FFDM. This compared with more substantial reductions in favour of DBT + FFDM in Lourenco 2014 of −28.7 (95% CI: −35.1 to −22.2) per 1000 screens and in Destounis 2014 [23] of −74.4 (95% CI: −105.6 to −43.1) per 1000 screens. All differences were highly statistically significant.

In all three studies a higher recall rate was observed in the FFDM group, but the magnitude of the differences varied across studies. Reduction in recall in the Friedewald 2014 [25–31] study was −16.2 (95% CI: −18.0 to −14.5) per 1000 screens, compared to a difference of −29.4 (95% CI: −36.0 to −22.8) in Lourenco 2014 [24] and a difference of −72.5 (95% CI: −104.7 to −40.2) in Destounis 2014 [23]. Due to the differences in the recall rate the results of the studies were not combined using meta-analysis (see Fig. 3).

All of the US studies were based on retrospective analysis. A paired analysis [31] for a subset of the women participating in Friedewald 2014 yielded results that were quantitatively similar to those observed in the parent study [25–31]. Reductions in recall and false positive rates were observed using DBT + FFDM: difference −27.6 (95% CI: −30.8 to −24.5) per 1000 screens and −29.5 (95% CI: −32.9 to −26.4) per 1000 screens, respectively. A higher cancer detection rate was also observed using DBT + FFDM: difference of 1.9 (95% CI: 1.2 to 2.9) per 1000 screens.

Discussion

Overview of findings

This systematic review identified five studies comparing DBT + FFDM with FFDM. Studies varied substantially as they were performed in different health systems with different screening paradigms. The five studies reported the relative cancer detection, false positive and recall rates for DBT + FFDM and FFDM. However, only limited evidence on interval cancers from follow-up is available at this time, therefore absolute sensitivity and specificity cannot be fully evaluated. To reflect significant differences in practice, analyses were conducted separately for European and US studies.

European studies

Two European studies observed higher cancer detection and invasive cancer detection rates using DBT + FFDM than FFDM alone. The differences were statistically significant within studies and in the pooled analysis.

Results for recall and false positive rates vary according to the double reading algorithm adopted. In STORM [19–22], where women were recalled if either reader reported a positive finding, both false positives and recall were lower using DBT + FFDM than using FFDM alone. In OTST [9,17,18] pre-arbitration false positive rates of individual readers, which are reflective of what would likely be found in a single reader paradigm, were lower for DBT + FDDM. Post-arbitration, higher recall and false positive rates were observed for DBT + FFDM compared to FFDM. However, DBT images were available at the arbitration meeting for both the FFDM and DBT + FFDM arms. This biases the results in favour of FFDM and suggests that the recall rate in the FFDM arm was underestimated

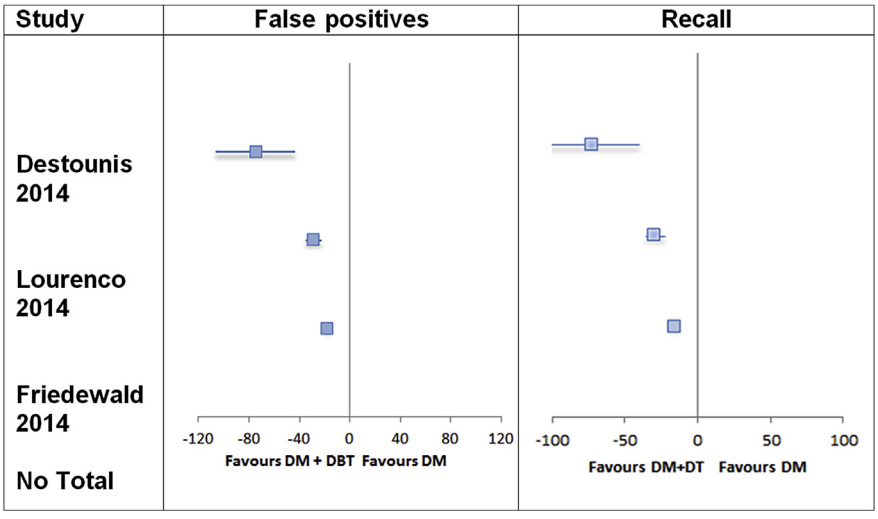

Fig. 3. Differences in false positive rates and recall rates per 1000 screens (US studies).

as an unknown number of cases in this arm may have been dismissed during arbitration based on DBT information. Despite this bias, the higher number of cancers detected with DBT + FFDM resulted in the positive predictive value of DBT + FFDM being similar to that of FFDM alone. Evidence from the European studies is currently insufficient to establish the exact impact of DBT on recall and false positives after consensus reading due to the lack of prospective blinded consensus reading in both large studies.

US studies

The US study cancer detection results were similar to those of the European studies, with two of the three US studies demonstrating an increased cancer detection rate. This included the largest study in the review, Friedewald 2014 [25–31], which analysed more than 450,000 examinations. The observed increase in all cancers detected and invasive cancers detected was, however, smaller than that observed in the European studies. This may be due to the shorter screening interval in the US (1 year) compared to Europe (2 years), the use of double reading in the European trials, or the relatively older age of women participating in European screening programmes. Although one small US study with an unpaired design observed a lower cancer detection rate using DBT + FFDM compared to FFDM, the difference was statistically insignificant.

The results of the US studies with regard to recall and false positive rates were much more consistent and showed sizable and statistically significant reductions in both recalls and false positives. This is consistent with the results of the European STORM study [19–22] and the pre-arbitration results of the OTST study which reflect what might be found with a single reader paradigm.

Limitations of the available evidence

While the results of this review suggest DBT is a promising technology, there are limitations to the available evidence. Data concerning fully blinded arbitration consensus are still lacking. Currently, there are only limited data on interval cancers and, hence, a comparative analysis of programme sensitivity is not possible. However, higher cancer detection rates with comparable or improved positive predictive value (PPV) were observed in studies with fully paired datasets (OTST [9,17,18], STORM [19–22] and the Rose et al. [31] analysis of the large US multicentre study). These results demonstrate the better relative sensitivity of DBT + FFDM compared to FFDM alone. The current limitation is that without complete follow-up data, the exact sensitivity and specificity of DBT + FFDM is not known. Studies reporting interval cancer data are expected to be reporting soon, at which point more information regarding exact sensitivities and specificities will be better understood.

The vast majority of data, including all studies reported in this review, have been collected using equipment marketed by a single vendor, Hologic Inc. Studies are ongoing using DBT systems developed by other manufacturers. The interim results of the Malmö Breast Tomosynthesis Screening Trial [32], which used the equipment of another vendor, was published in May 2015, almost six months after the deadline of the search criteria for this review (11/2014). This trial studied single view breast tomosynthesis in 7500 women and reported an increase in cancer detection by 43% and recall with maintained PPV, with DBT compared to FFDM. While the results of this study appear comparable to the evidence included in this review, caution must be used when extrapolating results from one system to another. As the methods of image acquisition and reconstruction can differ significantly between systems, clinical performance may also differ.

Parallel to this systematic review, Lauby-Secretan [6] published the Viewpoint of the IARC Working Group on breast cancer screening, which was based on multiple systematic searches and which also includes conclusions on the results of DBT. Using different methodology and a different data base, the conclusions by Lauby-Secretan and this systematic review are quite similar. This fact supports the robustness of the results and interpretations. Both conclusions are, however, limited by the availability of data at the time of the searches.

Implications for practice

Overall, the evidence suggests that cancer detection rates and invasive cancer detection rates are higher using DBT + FFDM than with FFDM, but non-invasive cancer detection rates are unchanged. Therefore, the addition of DBT to screening programmes has the potential to reduce morbidity and mortality associated with breast cancer by increasing early detection of tumours.

Evidence suggests that recall and false positive rates may be lower using DBT + FFDM, especially for single reader paradigms such as those common in the US. These reductions are particularly notable given recent emphasis on the potential harms of false positives and would reduce the number of women experiencing anxiety and distress caused by false positive examinations [33]. Furthermore, reductions in recalls have the clear added benefit of decreasing programme costs, particularly in the US where recall rates are higher than typically reported in Europe. With fewer women recalled for false positive findings, fewer diagnostic mammography, breast ultrasound and biopsies are performed while cancer detection is maintained.

Definitive conclusions regarding the impact of DBT on overdiagnosis cannot be made based on current data. However, overdiagnosis may be associated with ductal carcinoma in situ (DCIS) and DBT has not been shown to increase the rate of non-invasive cancer detection. Current data concerning the incremental invasive cancers detected by DBT do not yet allow final conclusions. However, small and low grade invasive carcinomas can mostly be treated avoiding aggressive treatments or overtreatment.

The adoption of DBT into screening programmes should also consider radiation exposure. Acquiring separate DBT and FFDM acquisitions results in approximately double the dose of a single FFDM acquisition. Replacing the additional DM acquisition by using synthesized views from the DBT dataset is now feasible and therefore reduces this concern. Evidence on the relative performance of DBT + synthesized DM is growing [9,34], but further research is needed.

Conclusion

Evidence from large scale studies in the US and Europe show that DBT + FFDM, compared to FFDM, yields higher invasive cancer detection rates, increasing the effectiveness of breast cancer screening. The use of DBT may reduce recalls and thereby reduce both programme costs and distress caused by a false negative recall.

Author notes

YHEC was commissioned to conduct the scoping review with funding provided by Hologic Ltd.

Conflict of interest statement

Within the past two years JS has been an employee of Novartis Healthcare Pvt. Ltd., India (this employment ceased prior to commencing work on this review). Within the past two years SHK

has received research funding from Siemens Healthcare, as well as having travel expenses paid for delivery of workshops organized by Siemens Healthcare. Siemens Healthcare had no influence on the content of SHK's contributions to the workshops.

Acknowledgements

The authors would like to thank Scott Pohlman and Chris Bartlett for their assistance in reviewing the manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.breast.2016.01.002.

References

[1] Ferlay J, Soerjomataram I, Ervik M, et al. GLOBOCAN 2012 v1.0, cancer incidence and mortality worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013.
[2] Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:E359–86.
[3] Shulman LN, Willett W, Sievers A, et al. Breast cancer in developing countries: opportunities for improved survival. J Oncol 2010;2010:595167.
[4] United States Preventive Services Task Force. Preventive services task force recommendation statement. Ann Intern Med 2009;151:716–26.
[5] Gotzsche PC, Jorgensen KJ. Screening for breast cancer with mammography. Cochrane Database Syst Rev 2013. Article Number: CD001877.
[6] Lauby-Secretan B, Scoccianti C, Loomis D, et al. Breast-cancer screening–viewpoint of the IARC Working Group. N Engl J Med 2015;372:2353–8.
[7] Hambly NM, McNicholas MM, Phelan N, et al. Comparison of digital mammography and screen-film mammography in breast cancer screening: a review in the Irish breast screening program. AJR Am J Roentgenol 2009;193:1010–8.
[8] Vinnicombe S, Pinto Pereira SM, McCormack VA, et al. Full-field digital versus screen-film mammography: comparison within the UK breast screening program and systematic review of published data. Radiology 2009;251:347–58.
[9] Skaane P, Bandos AI, Eben EB, et al. Two-view digital breast tomosynthesis screening with synthetically reconstructed projection images: comparison with digital breast tomosynthesis with full-field digital mammographic images. Radiology 2014;271:655–63.
[10] Lei J, Yang P, Zhang L, et al. Diagnostic accuracy of digital breast tomosynthesis versus digital mammography for benign and malignant lesions in breasts: a meta-analysis. Eur Radiol 2014;24:595–602.
[11] Houssami N, Skaane P. Overview of the evidence on digital breast tomosynthesis in breast cancer detection. Breast 2013;22:101–8.
[12] Houssami N. Digital breast tomosynthesis (3D-mammography) screening: data and implications for population screening. Expert Rev Med Devices 2015;12:377–9.
[13] DTA Editorial Team: Cochrane handbook for DTA reviews. London: The Cochrane Collaboration; 2013.
[14] Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions (ed 5.1.0 [updated March 2011]). London: The Cochrane Collaboration; 2011.
[15] Whiting PF, Weswood ME, Rutjes AW, et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol 2006;6:9.
[16] Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529–36.
[17] Skaane P, Bandos AI, Gullien R, et al. Prospective trial comparing full-field digital mammography (FFDM) versus combined FFDM and tomosynthesis in a population-based screening programme using independent double reading with arbitration. Eur Radiol 2013;23:2061–71.
[18] Skaane P, Bandos AI, Gullien R, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. Radiology 2013;267:47–56.
[19] Bernardi D, Caumo F, Macaskill P, et al. Effect of integrating 3D-mammography (digital breast tomosynthesis) with 2D-mammography on radiologists' true-positive and false-positive detection in a population breast screening trial. Eur J Cancer 2014;50:1232–8.
[20] Caumo F, Bernardi D, Ciatto S, et al. Incremental effect from integrating 3D-mammography (tomosynthesis) with 2D-mammography: increased breast cancer detection evident for screening centres in a population-based trial. Breast 2014;23:76–80.
[21] Ciatto S, Houssami N, Bernardi D, et al. Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. Lancet Oncol 2013;14:583–9.
[22] Houssami N, Macaskill P, Bernardi D, et al. Breast screening using 2D-mammography or integrating digital breast tomosynthesis (3D-mammography) for single-reading or double-reading - evidence to guide future screening strategies. Eur J Cancer 2014;50:1799–807.
[23] Destounis S, Arieno A, Morgan R. Initial experience with combination digital breast tomosynthesis plus full field digital mammography or full field digital mammography alone in the screening environment. J Clin Imaging Sci 2014;4:9.
[24] Lourenco AP, Barry-Brooks M, Baird G, et al. Changes in recall type and patient treatment following implementation of screening digital breast tomosynthesis. Radiology 2014:140317. Sep 22: 140317 [Epub ahead of print].
[25] Durand MA, Haas BM, Yao XP, et al. Early clinical experience with digital breast tomosynthesis for screening mammography. Radiology 2015;274:85–92.
[26] Friedewald SM, Rafferty EA, Rose SL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. J Am Med Assoc 2014;311:2499–507.
[27] Greenberg JS, Javitt MC, Katzen J, et al. Clinical performance metrics of 3D digital breast tomosynthesis compared with 2D digital mammography for breast cancer screening in community practice. AJR Am J Roentgenol 2014;203:687–93.
[28] Haas BM, Kalra V, Geisel J, et al. Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening. Radiology 2013;269:694–700.
[29] McCarthy AM, Kontos D, Synnestvedt M, et al. Screening outcomes following implementation of digital breast tomosynthesis in a general-population screening program. J Natl Cancer Inst 2014;106:dju316.
[30] Rose SL, Tidwell AL, Bujnoch LJ, et al. Implementation of breast tomosynthesis in a routine screening practice: an observational study. AJR AM J Roentgenol 2013;200:1401–8.
[31] Rose SL, Tidwell AL, Ice MF, et al. A reader study comparing prospective tomosynthesis interpretations with retrospective readings of the corresponding FFDM examinations. Acad Radiol 2014;21:1204–10.
[32] Lang K, Andersson I, Rosso A, et al. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the malmo breast tomosynthesis screening trial, a population-based study. Eur Radiol 2015;26:184–90.
[33] Welch HG, Passow HJ. Quantifying the benefits and harms of screening mammography. JAMA Intern Med 2014;174:448–54.
[34] Zuley ML, Bandos AI, Ganott MA, et al. Digital breast tomosynthesis versus supplemental diagnostic mammographic views for evaluation of noncalcified breast lesions. Radiology 2013;266:89–95.