



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

Daniela Bernardi, Petra Macaskill, Marco Pellegrini, Marvi Valentini, Carmine Fantò, Livio Ostillo, Paolina Tuttobene, Andrea Luparia, Nehmat Houssami

Summary

Background Breast tomosynthesis (pseudo-3D mammography) improves breast cancer detection when added to 2D mammography. In this study, we examined whether integrating 3D mammography with either standard 2D mammography acquisitions or with synthetic 2D images (reconstructed from 3D mammography) would detect more cases of breast cancer than 2D mammography alone, to potentially reduce the radiation burden from the combination of 2D plus 3D acquisitions.

Methods The Screening with Tomosynthesis Or standard Mammography-2 (STORM-2) study was a prospective population-based screening study comparing integrated 3D mammography (dual-acquisition 2D–3D mammography or 2D synthetic–3D mammography) with 2D mammography alone. Asymptomatic women aged 49 years or older who attended population-based screening in Trento, Italy were recruited for the study. All participants underwent digital mammography with 2D and 3D mammography acquisitions, with the use of software that allowed synthetic 2D mammographic images to be reconstructed from 3D acquisitions. Mammography screen-reading was done in two parallel double-readings conducted sequentially for 2D acquisitions followed by integrated acquisitions. Recall based on a positive mammography result was defined as recall at any screen read. Primary outcome measures were a comparison between integrated (2D–3D or 2D synthetic–3D) mammography and 2D mammography alone of the number of cases of screen-detected breast cancer, the cancer detection rate per 1000 screens, the incremental cancer detection rate, and the number and percentage of false-positive recalls.

Findings Between May 31, 2013, and May 29, 2015, 10 255 women were invited to participate, of whom 9672 agreed to participate and were screened. In these 9672 participants (median age 58 years [IQR 53–63]), screening detected 90 cases of breast cancer, including 74 invasive breast cancers, in 85 women (five women had bilateral breast cancer). To account for these bilateral cancers in cancer detection rate estimates, the number of screens used for analysis was 9677. Both 2D–3D mammography (cancer detection rate 8.5 per 1000 screens [82 cancers detected in 9677 screens]; 95% CI 6.7–10.5) and 2D synthetic–3D mammography (8.8 per 1000 [85 in 9677]; 7.0–10.8) had significantly higher rates of breast cancer detection than 2D mammography alone (6.3 per 1000 [61 in 9677], 4.8–8.1; $p < 0.0001$ for both comparisons). The cancer detection rate did not differ significantly between 2D–3D mammography and 2D synthetic–3D mammography ($p = 0.58$). Compared with 2D mammography alone, the incremental cancer detection rate from 2D–3D mammography was 2.2 per 1000 screens (95% CI 1.2–3.3) and that from 2D synthetic–3D mammography was 2.5 per 1000 (1.4–3.8). Compared with the proportion of false-positive recalls from 2D mammography alone (328 of 9587 participants not found to have cancer at assessment) [3.42%; 95% CI 3.07–3.80], false-positive recall was significantly higher for 2D–3D mammography (381 of 9587 [3.97%; 3.59–4.38], $p = 0.00063$) and for 2D synthetic–3D mammography (427 of 9587 [4.45%; 4.05–4.89], $p < 0.0001$).

Interpretation Integration of 3D mammography (2D–3D or 2D synthetic–3D) detected more cases of breast cancer than 2D mammography alone, but increased the percentage of false-positive recalls in sequential screen-reading. These results should be considered in the context of the trade-off between benefits and harms inherent in population breast cancer screening, including that significantly increased breast cancer detection from integrating 3D mammography into screening has the potential to augment screening benefit and also possibly contribute to overdiagnosis.

Funding None.

Introduction

The Screening with Tomosynthesis Or standard Mammography (STORM) trial,^{1,2} and the Oslo tomosynthesis trial,^{3,4} both of which were implemented

prospectively within population-based screening programmes, provided evidence that tomosynthesis (pseudo-3D mammography) combined with 2D mammography improved breast cancer detection

Lancet Oncol 2016

Published Online

June 23, 2016

[http://dx.doi.org/10.1016/S1470-2045\(16\)30101-2](http://dx.doi.org/10.1016/S1470-2045(16)30101-2)

See Online/Comment

[http://dx.doi.org/10.1016/S1470-2045\(16\)30155-3](http://dx.doi.org/10.1016/S1470-2045(16)30155-3)

U.O. Senologia Clinica e Screening Mammografico, Department of Diagnostics, Ospedale di Trento, Azienda Provinciale Servizi Sanitari, Trento, Italy (D Bernardi MD, M Pellegrini MD, M Valentini MD, C Fantò MD, L Ostillo MD, P Tuttobene MD, A Luparia MD); and School of Public Health, Sydney Medical School, University of Sydney, Sydney, NSW, Australia

(Prof P Macaskill PhD, Prof N Houssami PhD)

Correspondence to: Prof Nehmat Houssami, School of Public Health (A27), Sydney Medical School, University of Sydney, Sydney 2006, NSW, Australia
nehmat.houssami@sydney.edu.au

Research in context

Evidence before this study

One author (NH) updated a previously reported literature search using the methods detailed in a systematic review from Houssami and Skaane. The search consisted of a MEDLINE search (search terms "breast neoplasm", combined with "tomosyn\$" or "3D-mammography" in title) in February, 2016, and contact with content experts. The search aimed to identify prospective screening studies comparing 2D mammography with 3D mammography and synthetic 2D mammography, or with 3D mammography alone. Two studies, both reporting interim analyses, provided relevant information: the Oslo tomosynthesis trial's interim analysis reported that radiologists interpreting tomosynthesis with synthetic 2D images had similar breast cancer detection rates as those interpreting dual-acquisition 2D and 3D mammography and helped inform our research plan. The Malmö Breast Tomosynthesis Screening Trial reported that standalone 3D mammography detected more breast cancers than 2D mammography. These studies reported heterogeneous results for false-positive recall from 3D mammography, partly caused by differences in analytical approaches and whether single-reading or double-reading was reported.

Added value of this study

Our work provides evidence from a prospective screening study comparing 3D mammography screening strategies

(2D–3D mammography and 2D synthetic–3D mammography) with standard 2D mammography using the same screening examinations and based on double-reading. The results show that both integrated 2D–3D mammography and 2D synthetic–3D mammography screening detected more breast cancers than 2D mammography alone; these integrated 3D screen-reading strategies have similar breast cancer detection rates. However, our study was not powered to detect small differences in cancer detection rate between 2D–3D and 2D synthetic–3D mammography. Screen reading using 3D mammography, when read sequentially after 2D mammography, had more false positives than 2D mammography alone; therefore, future research should explore other approaches, such as upfront interpretation of 2D synthetic–3D or 3D-only screening.

Implications of all the available evidence

The results of this study should be factored into policy decisions, can be used to inform women regarding 3D mammography screening, and would support further evaluation of 2D–3D versus 2D synthetic–3D mammography in larger studies embedded in screening services.

compared with 2D mammography alone. Several retrospective studies^{5–9} have subsequently shown that integrated 2D–3D mammography improves screen detection measures compared with 2D mammography;^{5–10} however, combined 2D–3D mammography entails acquisition of both 2D and 3D images, which roughly doubles the amount of radiation delivered to the breast.^{11–13} An interim analysis from the Oslo study⁴ showed that 3D mammography acquisitions from which synthetic 2D images are reconstructed yield breast cancer detection frequencies that are similar to dual acquisition of 2D and 3D images; however, so far, no prospective screening trials of 3D mammography with synthetic 2D images have reported final results. We postulated that integration of 3D mammography with either acquired (standard) 2D images or with reconstructed images (2D synthetic) would similarly detect significantly more cases of breast cancer than 2D mammography alone. If this hypothesis was proven to be correct, then future recommendations regarding the adoption of integrated 2D and 3D mammography could eliminate dual acquisitions by using 2D synthetic rather than acquired 2D mammography.

In the STORM-2 prospective population-based screening trial, we compared tomosynthesis (3D mammography) with standard 2D digital mammography using two screening strategies. Our primary aims were to assess whether 3D mammography with synthetic images (2D synthetic–3D mammography) or with dual-acquisition

2D–3D mammography would detect significantly more cancers at screening than standard 2D mammography alone, and to compare breast cancer detection from these integrated screening strategies within the same participants to inform future breast screening practice.

Methods

Study design and participants

STORM-2 is a prospective population-based screening study comparing tomosynthesis (3D mammography) with standard 2D digital mammography. Women aged 49 years or older who attended population-based screening through the Trento screening programme (Trento Centre, Trento, Italy) were invited to have screening with both 2D and 3D mammography acquisitions. Participants might have already had 3D mammography screening in our previous trial (STORM) in 2011–12.¹² Participants were asymptomatic women attending for biennial screening mammography provided to women at standard (population) risk for breast cancer.

The study was granted institutional ethics approval by the local health district ethics committee for clinical research, and written informed consent was obtained from all screening participants before screening.

Procedures

This study used two screening strategies: one integrating 2D mammography with 3D mammography (2D and 3D acquisitions) and another using 3D mammography

acquisitions enabling 2D image reconstruction (3D mammography with 2D synthetic images). All participants underwent digital mammography with 2D and 3D mammography acquisitions, with the use of software that allowed synthetic 2D mammographic images to be reconstructed from 3D acquisitions (Selenia Dimensions Unit operated in COMBO mode with C-View 2D-software [Hologic Inc, Bedford, MA, USA]). Both 2D and 3D images were acquired at the same screening examination with a single breast positioning and compression per view, with mediolateral oblique and cranio-caudal views obtained for 2D and 3D acquisitions. Women who declined to participate received standard 2D mammography.

Mammography screen-reading was done sequentially in two parallel double-readings yielding paired data for each screening examination in each screen-reading strategy. Figure 1 shows the two independently reported double-reading strategies: in one double-reading strategy (reads A and B) screens were interpreted sequentially by radiologists viewing standard 2D mammography alone and then re-interpreted by the same radiologists on the same day using integrated 2D–3D mammography; and in the second double-reading strategy (reads C and D), screens were interpreted sequentially by a different pair of readers, initially using 2D synthetic mammography (reconstructed from 3D acquisition) and then re-interpreted by the same radiologists on the same day using integrated 2D synthetic–3D mammography. Thus, integrated screen-reading was based on joint interpretation of 2D–3D mammography or 2D synthetic with 3D mammography, and does not refer to analytical combinations. Radiologists were required to report each screening mammogram according to this sequence, and to record whether to recall women at each screen-reading phase before progressing to the next phase. All screen reads (A/B and C/D) were based on the same screening examinations. For each screen, breast density data were also recorded.

Seven dedicated breast radiologists participated in screen-reading and had a mean of 13 years (range 3–23) of experience in breast imaging; all had received training in 3D mammography and had been using 3D mammography for a mean of 2.7 years (range 2–3). Six of the seven radiologists participated in both double-readings, whereas one radiologist reported in the 2D synthetic–3D double-readings only (reads C and D) because of work scheduling constraints. When previous screening mammograms were available, these were displayed at the time of screen reading as standard practice. Breast density was reported at 2D screen reading using Breast Imaging Reporting and Data System (BI-RADS)¹⁴ classification, with density from all reads derived from the majority score (or with use of arbitration when necessary) and analysed in two groups: 1–2 (less dense) and 3–4 (more dense).^{1,14}

Mammography interpretation was based on independent double-reading for each double-reading

strategy, given that double-reading is done in organised screening programmes. A screen was classified as positive and the woman recalled to assessment for further investigations if recalled by either screen reader in either of the double-reading strategies based on recall at any screen-reading phase.

Outcomes

Our primary outcome measures were the number of detected cancers and the cancer detection rate per 1000 screens, the incremental cancer detection rate attributable to integrated 2D–3D or 2D synthetic–3D mammography, and the number and percentage of false-positive recalls. The incremental cancer detection rate estimates the additional cancers detected per 1000 screens for screen-reading that integrates 2D and 3D mammography compared with 2D mammography alone. Outcomes were ascertained on the basis of excision histology in those who received surgery, or based on the completed assessment outcome, inclusive of work-up imaging (with or without histology from core needle biopsy), in recalled participants. Hence, outcomes are based on detection at screening and assessment in recalled women.

Statistical analysis

Our sample size was planned to detect a difference in cancer detection rate of two per 1000 screens (in each

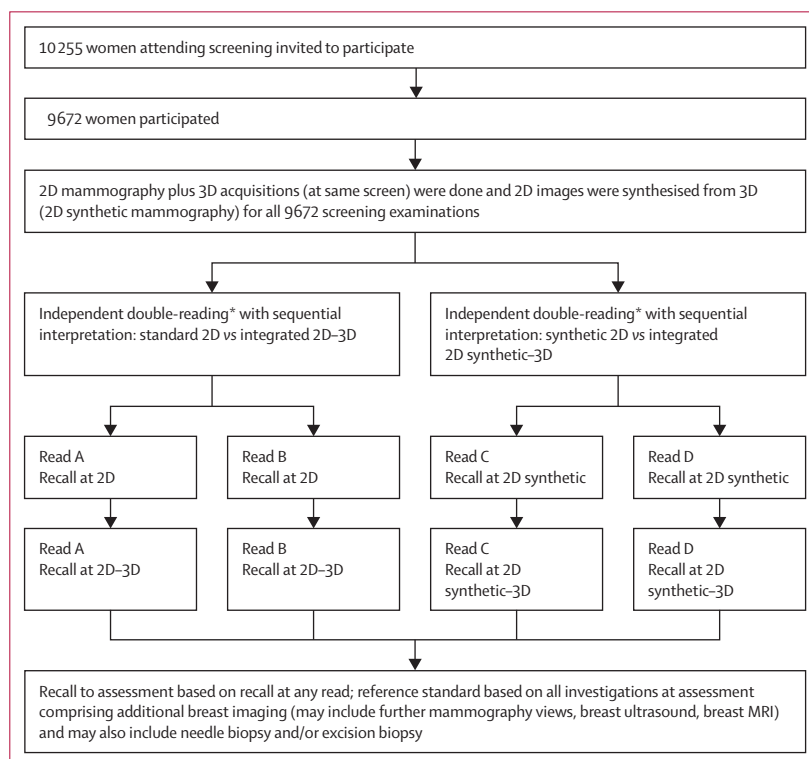


Figure 1: Study profile

*Same screens for each participant were used in all screen-reads.

double-reading strategy) between 2D mammography and integrated 2D–3D mammography, based on the assumption that 2D mammography alone would detect five cases of breast cancer per 1000 screens and that screen-reading integrating 3D mammography (2D synthetic–3D mammography or 2D–3D mammography) would detect seven cases of breast cancer per 1000 screens, approximated from STORM.¹ Because most of the screens in the study represented incident (repeat) screening, we assumed an underlying breast cancer prevalence of 0.5% for 2D mammography screening and assumed dependence between sequential readings as informed by our earlier study.¹ Based on these assumptions and using methods¹⁵ for ascertaining sample size for studies using McNemar's test¹⁶ for paired binary data, we estimated that a sample size of 7850 screens for 80% power, or 10 500 screens for 90% power, would be needed to detect differences in cancer detection rates between screening strategies.

We used two-way tables to compare the number of cancers detected exclusively at each screen-reading phase per double-reading strategy (2D vs 2D–3D and 2D

synthetic vs 2D synthetic–3D), and across double-reading strategies for relevant comparisons (2D vs 2D synthetic–3D and 2D–3D vs 2D synthetic–3D); we did the same analyses for false-positive recall. In our primary analyses we calculated and compared cancer detection rates per 1000 screens at 2D–3D or 2D synthetic–3D screening versus 2D mammography alone given that 2D mammography represents the present standard method of breast screening; we also compared cancer detection rates for 2D–3D mammography and 2D synthetic–3D mammography. We calculated the incremental cancer detection rate attributable to integrated 2D–3D or 2D synthetic–3D screening versus 2D mammography. We calculated the false-positive recall frequency, the proportion of false-positive recalls per double-reading strategy, and compared the proportion of false-positive recalls for 2D–3D and for 2D synthetic–3D mammography with that from 2D mammography alone.

We stratified our main analyses by age and breast density. We also planned to explore stratification by prevalent or incident screening; however, most screens (9058/9672 [$>93\%$]) were incident (repeat) screens and therefore this stratification was not informative. SAS/STAT version 9.4 was used for analyses, with the use of exact methods to compute 95% CIs for cancer detection rate and percentage false-positive recall estimates, and p values for the McNemar test.¹⁶ We used StatsDirect version 3.0.167 to compute exact CIs for differences in paired proportions.

Role of the funding source

There was no funding source for this study. DB and NH had full access to all the data and had final responsibility for the decision to submit for publication.

Results

Between May 31, 2013, and May 29, 2015, 10 255 women were invited to participate in STORM-2, of whom 9672 (94%) agreed to participate and were screened. The median age of the participants was 58 years (IQR 53–63). The estimated mean glandular dose of radiation per view was 1.36 mGy (SD 0.51) from 2D mammography, 1.87 mGy (0.67) from 3D mammography, and 3.22 mGy (1.16) from dual-acquisition mammography (2D plus 3D). Screening detected 90 breast cancers in 85 women (bilateral breast cancer was detected in five women); 74 were invasive breast cancers (64 ductal and ten lobular) and 16 were ductal carcinoma in situ. To account for these bilateral cancers in cancer detection rate estimates, the number of screens used for analysis was 9677. Table 1 shows the characteristics of these cancers according to detection method. The mean tumour size for invasive cancers was 12.7 mm (SD 7.8) for the 46 invasive cancers detected by 2D mammography and also by integrated 2D–3D or 2D synthetic–3D screening, and 11.6 mm (9.4) for the 28 invasive cancers detected only by 2D–3D or 2D synthetic–3D screening.

	Cancers detected at standard 2D mammography and also at integrated 2D–3D or 2D synthetic–3D screening (n=61 cancers)	Cancers detected only at integrated 2D–3D or 2D synthetic–3D screening (n=29 cancers)
Pathological tumour size (pT) category		
pTis (ductal carcinoma in situ)	15 (25%)	1 (4%)
pT1a (≤ 5 mm)	7 (11%)	5 (17%)
pT1b (>5 mm– ≤ 10 mm)	15 (25%)	9 (31%)
pT1c (>10 mm– ≤ 20 mm)	18 (30%)	11 (38%)
pT2 (>20 mm– ≤ 50 mm)	6 (10%)	3 (10%)*
Nodal status		
Negative for metastases	43 (71%)	25 (86%)
Positive for metastases	7 (11%)	3 (10%)†
NA	11 (18%)	1 (3%)
Tumour grade		
Grade 1	20 (33%)	9 (31%)
Grade 2	28 (46%)	16 (55%)
Grade 3	13 (21%)	4 (14%)
ER/PR status‡		
Positive	45 (74%)	26 (90%)
Negative	3 (5%)	1 (3%)
NA	12 (20%)	2 (7%)
HER2 status		
Positive	4 (7%)	1 (3%)
Negative	42 (69%)	25 (86%)
NA	15 (25%)	3 (10%)

Data are n (%) or n. NA=not performed, not applicable, or data not available. ER=oestrogen receptor. PR=progesterone receptor. *Includes one pT3. †Includes one case with micro-metastases. ‡Positive if either ER or PR or both are positive.

Table 1: Characteristics of breast cancers detected

Figure 2 presents two-way paired cancer detection data for the 90 screen-detected breast cancers. Cancer detection rates for 2D mammography, 2D–3D mammography, and 2D synthetic–3D mammography are summarised in table 2. Both integrated 2D–3D mammography and 2D synthetic–3D mammography screening had significantly higher cancer detection rates than 2D mammography alone (table 2). Cancer detection rates (and therefore also the related incremental cancer detection rates) did not differ significantly between 2D–3D mammography and 2D synthetic–3D mammography ($p=0.58$).

In 1771 (18%) of 9672 participants previously screened with 2D–3D mammography in our earlier study (STORM),^{1,2} estimated cancer detection rates were 6.8 per 1000 screens for 2D mammography alone (12 cancers in 1771 screens; 95% CI 3.5–11.8), 8.5 per 1000 for 2D–3D mammography (15 in 1771; 4.7–13.9), and 8.5 per 1000 for 2D synthetic–3D mammography (15 in 1771; 4.7–13.9).

Integrated 2D–3D mammography and 2D synthetic–3D mammography resulted in incremental cancer detection compared with 2D mammography across age group and density categories, although the cancer detection rate for 2D synthetic–3D mammography did not differ significantly from that for 2D mammography

in women aged 60 years and older (table 2). Incremental detection from integrated screen-reading was more frequent in women younger than 60 years compared with those older than 60 and in those with denser breasts (table 2). Cancer detection rates (and therefore also incremental cancer detection rates) did not differ between 2D–3D mammography and 2D synthetic–3D mammography in stratified analyses (all $p>0.20$; appendix).

Figure 3 summarises two-way paired data tables for women classified as not having breast cancer ($n=9587$), including those with false-positive recalls, for all double-reading strategies. False-positive recalls, counting both double-readings (A/B and C/D), occurred in 746 of 9587 participants (7.78%; 95% CI 7.25–8.34); the proportion of false-positive recalls did not differ significantly between sequential double-reading of 2D and 2D–3D mammography and sequential double-reading of 2D synthetic and 2D synthetic–3D mammography (table 3). Compared with false-positive recalls from 2D mammography alone, the proportion of false-positive recalls for integrated 2D–3D mammography and 2D synthetic–3D mammography were significantly higher (table 3). Estimated false-positive recall differed between integrated screen-reading strategies (0.48%; 95% CI 0.05–0.91, $p=0.033$; table 3).

See Online for appendix

A Breast cancers detected at standard 2D mammography vs 2D–3D mammography; double-reading A and B ($p<0.0001\ddagger$)			
	2D–3D positive	2D–3D negative	Total
2D positive	61	0	61
2D negative	21	8*	29
Total	82	8	90†

B Breast cancers detected at standard 2D mammography vs 2D synthetic–3D mammography; cross-comparison of double-read A and B for 2D alone and double-read C and D for 2D synthetic–3D mammography ($p<0.0001\ddagger$)			
	2D synthetic–3D positive	2D synthetic–3D negative	Total
2D positive	58	3	61
2D negative	27	2*	29
Total	85	5	90†

C Breast cancers detected at synthetic 2D mammography vs 2D synthetic–3D mammography; double-reading C and D ($p<0.0001\ddagger$)			
	2D synthetic–3D positive	2D synthetic–3D negative	Total
2D synthetic positive	66	0	66
2D synthetic negative	19	5*	24
Total	85	5	90†

D Breast cancers detected at integrated screening: 2D–3D mammography vs 2D synthetic–3D mammography; cross-comparison of double-read A and B and double-read C and D ($p=0.58\ddagger$)			
	2D synthetic–3D positive	2D synthetic–3D negative	Total
2D–3D positive	77	5	82
2D–3D negative	8	0*	8
Total	85	5	90†

Figure 2: Two-way paired tables for breast cancer detection based on double-reading strategies

(A) Standard 2D mammography vs 2D–3D mammography. (B) Standard 2D mammography vs 2D synthetic–3D mammography. (C) 2D synthetic mammography vs 2D synthetic–3D mammography. (D) 2D–3D mammography vs 2D synthetic–3D mammography. *Based on cancers detected in the study population and does not include data for interval cancers—this does not affect the comparative detection data shown in this cross-tabulation. †Total data for 90 breast cancers double-count the five women who were found to have bilateral cancers. ‡Exact p values for McNemar's test for paired binary data (1 degree of freedom).

	Detected cancers, n	CDR per 1000 screens (95% CI)	p value*	Incremental CDR per 1000 screens (95% CI) attributed to integrating 3D screening vs 2D alone†
All screening participants (n=9672, analysed as n=9677)‡				
Standard digital 2D mammography	61	6.3 (4.8–8.1)
Integrated 2D–3D mammography	82	8.5 (6.7–10.5)	<0.0001	2.2 (1.2–3.3)
Integrated 2D synthetic–3D mammography	85	8.8 (7.0–10.8)	<0.0001	2.5 (1.4–3.8)
Stratified by age				
<60 years (n=5745)				
Standard digital 2D mammography	21	3.7 (2.3–5.6)
Integrated 2D–3D mammography	36	6.3 (4.4–8.7)	<0.0001	2.6 (1.3–4.3)
Integrated 2D synthetic–3D mammography	40	7.0 (5.0–9.5)	<0.0001	3.3 (1.9–5.2)
≥60 years (n=3932)				
Standard digital 2D mammography	40	10.2 (7.3–13.8)
Integrated 2D–3D mammography	46	11.7 (8.6–15.6)	0.031	1.5 (0.03–3.3)
Integrated 2D synthetic–3D mammography	45	11.4 (8.4–15.3)	0.23	1.3 (–0.6–3.3)
Stratified by breast density				
Density 1–2 (less dense; n=7085)				
Standard digital 2D mammography	41	5.8 (4.2–7.8)
Integrated 2D–3D mammography	48	6.8 (5.0–9.0)	0.016	1.0 (0.1–2.0)
Integrated 2D synthetic–3D mammography	49	6.9 (5.1–9.1)	0.057	1.1 (0.0–2.4)
Density 3–4 (more dense; n=2592)				
Standard digital 2D mammography	20	7.7 (4.7–11.9)
Integrated 2D–3D mammography	34	13.1 (9.1–18.3)	<0.0001	5.4 (2.6–9.0)
Integrated 2D synthetic–3D mammography	36	13.9 (9.7–19.2)	<0.0001	6.2 (3.2–10.0)

CDR=cancer detection rate. *Exact p value for McNemar's test comparing paired binary data; p refers to comparison of CDR for screening integrating 3D mammography vs standard 2D mammography. †Incremental CDR estimates the additional cancers detected per 1000 screens for screen-reading that integrates 3D mammography (2D–3D or 2D synthetic–3D) compared with 2D mammography alone based on paired data. For example, in all screening participants, 85 cancers were detected by integrated 2D synthetic–3D mammography and 61 were detected by standard 2D mammography; 58 of these cancers were detected by both methods. These data allow construction of a two-way paired table (analogous to those shown in figure 2) to compute the McNemar test and estimated incremental CDR and corresponding 95% CI. ‡Estimated CDR for 9672 screens are analysed as 9677 to allow for the five participants (in both the numerator and the denominator) who were found to have bilateral cancers with different detection results between breasts.

Table 2: Breast cancer detection rates and incremental cancer detection from integrating 3D mammography in screen-reading versus 2D mammography alone

We also compared the false-positive recalls for integrated mammography versus 2D mammography stratified by age and breast density (table 3). In both subgroups, a higher proportion of false-positive recalls were noted for integrated 3D screen-readings than for 2D mammography alone, although the false-positive recalls for 2D–3D mammography did not differ significantly from that for 2D mammography in women younger than 60 years (table 3). Incremental false-positive recalls from integrated 3D screen readings were consistently higher for 2D synthetic–3D mammography than for 2D–3D mammography (table 3), differing significantly in women younger than 60 years of age ($p=0.023$) and in those with denser breasts ($p=0.039$; appendix).

Discussion

The results of the prospective STORM-2 trial show that either of the screen-reading strategies that used 3D mammography—dual-acquisition 2D–3D or 2D synthetic–3D mammography (from 3D-only acquisition)—detected significantly more cases of breast cancer than 2D mammography alone, yielding incremental cancer detection rates of 2.2 per 1000 screens and 2.5 per 1000 screens, respectively. Additionally, the types of breast cancers detected only when 3D mammography was integrated in screen-reads were similarly distributed compared with those detected by 2D mammography, except that almost all breast cancers detected only at 3D mammography were invasive—thus, there were relatively higher proportions of pT1a–c cancer in those detected only at 3D mammography compared with those detected by 2D mammography. We are aware that 2D mammography is the only breast imaging technique endorsed in existing screening recommendations,^{11,17} and that so far, whether the enhanced detection from 3D mammography confers incremental screening benefit in terms of breast cancer mortality outcomes is unknown.^{11,17,18} However, in the context of emerging adoption of 2D–3D mammography screening,^{1–3,5–10,18,19} we sought to address a major knowledge gap about whether integrated 2D–3D mammography could be potentially replaced by 3D mammography acquisitions that also synthesise 2D images (2D synthetic–3D mammography) to avoid the radiation burden from dual-acquisition mammography (2D plus 3D). The evidence from our work can be used to inform future screening practice and further studies using 2D synthetic–3D mammography.

We found that integrating 3D mammography in screen-reading improved breast cancer detection across age group and density strata; however, a large amount of the effect of 3D mammography was observed in women younger than 60 years, and in those with denser breasts. We cannot meaningfully compare our findings to those from other studies because none have reported a direct comparison of these integrated screening strategies for double-reading nor for completed (final) results, and none of the prospective tomosynthesis screening trials have yet reported age-stratified or density-stratified results for 2D synthetic–3D mammography.^{4,20} Our search of the published literature (which updated a literature review by Houssami and Skaane²¹) showed that there is little published evidence about this issue, and it is limited to one informative interim report from the Oslo study⁴ that showed similar breast cancer detection for 2D–3D mammography and 2D synthetic–3D mammography in single-reading analysis. The Malmö study's interim report²⁰ showed that 3D mammography detected significantly more breast cancers than 2D mammography but the study did not investigate 2D synthetic mammography. One retrospective study⁷ compared women screened with 2D mammography and those screened with dual-acquisition 2D–3D

A Standard 2D mammography vs 2D–3D mammography; double-reading A and B (*p=0.00063)			
	2D–3D positive	2D–3D negative	Total
2D positive	238	90	328
2D negative	143	9116	9259
Total	381	9206	9587

B Standard 2D mammography vs 2D synthetic–3D mammography; cross-comparison of double-read A and B for 2D mammography and double-read C and D for 2D synthetic–3D mammography (*p<0.0001)			
	2D synthetic–3D positive	2D synthetic–3D negative	Total
2D positive	134	194	328
2D negative	293	8966	9259
Total	427	9160	9587

C Synthetic 2D mammography vs 2D synthetic–3D mammography; double-reading C and D (*p<0.0001)			
	2D synthetic–3D positive	2D synthetic–3D negative	Total
2D synthetic positive	286	67	353
2D synthetic negative	141	9093	9234
Total	427	9160	9587

D Integrated screening: 2D–3D mammography vs 2D synthetic–3D mammography; cross-comparison of double-read A and B and double-read C and D (*p=0.033)			
	2D synthetic–3D positive	2D synthetic–3D negative	Total
2D–3D positive	181	200	381
2D–3D negative	246	8960	9206
Total	427	9160	9587

Figure 3: Two-way paired tables for screens classified as not having breast cancer, inclusive of false-positive recall, based on double-reading strategies

(A) Standard 2D mammography vs 2D–3D mammography. (B) Standard 2D mammography vs 2D synthetic–3D mammography. (C) 2D synthetic mammography vs 2D synthetic–3D mammography. (D) 2D–3D mammography vs 2D synthetic–3D mammography. *Exact p values for McNemar's test for paired binary data (1 degree of freedom).

mammography using a “before versus after” design, and showed that 2D–3D mammography was associated with significantly increased cancer detection rates compared with those for 2D mammography only in women younger than 50 years.

The STORM-2 trial has some limitations that should be considered. To support comparison against standard 2D mammography, and also between the two integrated strategies, our trial entailed sequential reads and duplication of the double-reading practice in organised screening programmes. This approach enabled us to address several knowledge gaps regarding tomosynthesis screening; however, it increased both breast cancer detection and false-positive recall across all screen-reads of the trial (and thus in both double-readings) because recall was based on recall from four sequential reads. We therefore present our results mainly per double-reading strategy in accord with screening practice and to inform population screening policy decisions; however, the sequential screen-readings might have affected the estimated false-positive recall proportions. Our false-positive recall of around 4–5% per double-reading strategy is within the acceptable level for screening,²² although we found that the addition of 3D mammography increased false-positive recall by modest but statistically significant increments relative to 2D mammography alone.

Data on the effect of 3D mammography on false-positive recall are heterogeneous and are confounded by both screen-reading approach (single/double and recall/arbitration methods) and study design: retrospective studies (which generally used single reading) have shown that 3D mammography reduces false-positive recall^{5–10} whereas prospective trials^{1,3,20} have shown that 3D mammography contributes around an additional 1% of false-positive recall in double-reading strategies. Our false-positive recall for integrated 2D–3D mammography (3.97%) was slightly higher than the estimated 3.7% recall rate from the Oslo trial's interim results for double-reading with arbitration.³ We think that sequential interpretation and recall based on either reader's recall (without arbitration for discordant reads), whereby recall was decided at either mammography modality by either reader, increased false-positive recall in our study. We anticipate that upfront integrated screen-reading with 2D–3D mammography or with 2D synthetic–3D mammography (without initial 2D-alone reading) and arbitration for discordant reads could yield lower false-positive recall than that observed in STORM-2. We also expect that increasing experience with 2D synthetic–3D mammography could further reduce false-positive recall given that STORM-2 represented the study radiologists' first screening experience using 2D

	False-positive recalls (n)	False-positive recalls, % (95% CI)	p value*	Difference in % false-positive recalls† (95% CI)
False-positive recall for each double-reading strategy based on all screens classified as not having breast cancer (n=9587)				
Double-reading A and B: recall for 2D or 2D–3D	471	4.91% (4.49–5.36)	0.34‡	0.24 (–0.24 to 0.71)‡
Double-reading C and D: recall for 2D synthetic or 2D synthetic–3D	494	5.15% (4.72–5.61)	..	
False-positive recall for standard 2D mammography vs integrated screening				
Standard digital 2D mammography	328	3.42% (3.07–3.80)
Integrated 2D–3D mammography	381	3.97% (3.59–4.38)	0.00063	0.55 (0.24 to 0.87)
Integrated 2D synthetic–3D mammography	427	4.45% (4.05–4.89)	<0.0001	1.03 (0.58 to 1.49)
Stratified by age				
<60 years (5704 screens)				
Standard digital 2D mammography	225	3.94% (3.45–4.48)
Integrated 2D–3D mammography	246	4.31% (3.80–4.87)	0.10	0.37 (–0.06 to 0.80)
Integrated 2D synthetic–3D mammography	286	5.01% (4.46–5.61)	0.00085	1.07 (0.45 to 1.70)
≥60 years (3883 screens)				
Standard digital 2D mammography	103	2.65% (2.17–3.21)
Integrated 2D–3D mammography	135	3.48% (2.92–4.10)	0.00054	0.82 (0.37 to 1.30)
Integrated 2D synthetic–3D mammography	141	3.63% (3.07–4.27)	0.0035	0.98 (0.34 to 1.64)
Stratified by breast density				
Density 1–2 (less dense; 7032 screens)				
Standard digital 2D mammography	227	3.23% (2.83–3.67)
Integrated 2D–3D mammography	253	3.60% (3.14–4.06)	0.051	0.37 (0.01 to 0.73)
Integrated 2D synthetic–3D mammography	272	3.87% (3.43–4.35)	0.014	0.64 (0.14 to 1.14)
Density 3–4 (more dense; 2555 screens)				
Standard digital 2D–mammography	101	3.95% (3.23–4.78)
Integrated 2D–3D mammography	128	5.01% (4.20–5.93)	0.0016	1.06 (0.42 to 1.73)
Integrated 2D synthetic–3D mammography	155	6.07% (5.17–7.06)	<0.0001	2.11 (1.13 to 3.12)

*Exact p value for McNemar's test for paired binary data; p value is for comparison of false-positive recall for screening integrating 3D mammography vs standard 2D mammography alone, except where otherwise specified.
†Estimated difference in % false-positive recalls compared with standard 2D mammography except where otherwise specified. ‡Denotes p value and difference in % false-positive recalls for comparison of double-readings shown in figure 1 (A/B vs C/D).

Table 3: False-positive recall proportion shown according to double-reading protocols, and comparison of false-positive recall for screening integrating 3D mammography vs standard 2D mammography alone

synthetic mammography: participating radiologists noted that synthetic 2D images showed enhanced visibility of parenchymal structures and also lesion details which contributed to more recall (including more cancer detection) than standard 2D mammograms.

Another situation that might have occurred in our trial is that some cancers might have been missed by all screen reads in the trial (these will be identified as interval cancers at follow-up); however, this does not affect estimates of the study's endpoint which is comparative breast cancer detection at screening.

A further limitation of our study is that it focused on screen detection measures, and its design and sample size cannot address long-term screening endpoints such as breast cancer mortality and overdiagnosis. These important outcomes need very large datasets and extended follow-up,¹⁸ and are beyond the scope of STORM-2. As noted in the US Preventive Services Task Force statement on breast screening, 2D synthetic–3D mammography has the potential to reduce radiation doses compared with 2D–3D mammography but little evidence exists for its screen-detection metrics.¹⁷ Although we do not advocate transitioning to 3D mammography screening based on STORM-2, we think that those recommending 2D–3D mammography should use our findings to inform women that 2D synthetic–3D mammography yields similar breast cancer detection as dual-acquisition 2D–3D mammography with the advantage of reducing radiation exposure by eliminating 2D acquisition; based on STORM-2, reduction in average glandular dose was around 42% compared with dual acquisition—a similar finding to that from the Oslo tomosynthesis trial's interim report.⁴ Although our study was not powered to detect very small differences in cancer detection rates between 2D–3D and 2D synthetic–3D mammography, radiation dose reduction from using 2D synthetic–3D mammography is a potential harm reduction to all screened women.

In conclusion, our prospective population screening study has shown that integrated 2D–3D or 2D synthetic–3D mammography provide similar breast cancer detection rates that are significantly higher than those for standard 2D mammography, in direct comparisons of these screen-reading strategies in the same participants. Integrated 2D–3D and 2D synthetic–3D screening also had higher rates of false-positive recall than standard 2D mammography, although that outcome might be mitigated through further experience and repeated screening with 3D mammography. Importantly, 2D synthetic–3D mammography reduces the radiation burden to the breast compared with 2D–3D mammography. The evidence from STORM-2 should be considered in the context of the trade-off between benefits and harms inherent in population breast cancer screening, including that significantly increased breast cancer detection from integrating 3D mammography into screening has the potential to augment screening benefit and also possibly contribute to overdiagnosis.

Contributors

DB contributed to study design and overall supervision of the study, data collection and interpretation, and writing and revisions of the report. PM contributed to statistical planning, data analysis and interpretation, and writing and revisions of the report. MP, MV, CF, LO, PT, and AL contributed to screen-reading, data collection, and review of the draft report. NH contributed to study design and scientific direction, data analysis and interpretation, literature search, and writing and revisions of the report.

Declaration of interests

DB has received travel support from Hologic. NH has received research support from the National Breast Council Foundation, Australia. All other authors declare no competing interests.

Acknowledgments

NH receives research support through a National Breast Cancer Foundation (NBCF, Australia) Breast Cancer Research Leadership Fellowship.

References

- 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**: 583–89.
- 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.
- 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.
- 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.
- Haas BM, Kalra V, Geisel J, Raghu M, Durand M, Philpotts LE. Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening. *Radiology* 2013; **269**: 694–700.
- Friedewald SM, Rafferty EA, Rose SL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. *JAMA* 2014; **311**: 2499–507.
- 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.
- Greenberg JS, Javitt MC, Katzen J, Michael S, Holland AE. 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.
- Durand MA, Haas BM, Yao X, et al. Early clinical experience with digital breast tomosynthesis for screening mammography. *Radiology* 2015; **274**: 85–92.
- Houssami N. Digital breast tomosynthesis (3D-mammography) screening: data and implications for population screening. *Expert Rev Med Devices* 2015; **12**: 377–79.
- Lauby-Secretan B, Scoccianti C, Loomis D, et al. Breast-cancer screening—viewpoint of the IARC Working Group. *N Engl J Med* 2015; **372**: 2353–58.
- Svahn TM, Houssami N, Sechopoulos I, Mattsson S. Review of radiation dose estimates in digital breast tomosynthesis relative to those in two-view full-field digital mammography. *Breast* 2015; **24**: 93–99.
- 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**: 1232–38.
- American College of Radiology (ACR). ACR BI-RADS: Breast imaging reporting and data system, Breast Imaging Atlas. Reston, VA: American College of Radiology, 2003.
- Lachenbruch PA. On the sample size for studies based on McNemar's test. *Stat Med* 1992; **11**: 1521–25.
- Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research, 4th edn. Oxford: Blackwell Science, 2002.
- Siu AL. US Preventive Services Task Force. Screening for Breast Cancer: US Preventive Services Task Force Recommendation Statement. *Ann Intern Med* 2016; **164**: 279–96.
- Houssami N, Miglioretti DL. Digital breast tomosynthesis. A brave new world of mammography screening. *JAMA Oncol* 2016; published online Feb 18. DOI:10.1001/jamaoncol.2015.5569.
- Houssami N, Lång K, Bernardi D, Tagliafico A, Zackrisson S, Skaane P. Digital breast tomosynthesis (3D-mammography) screening: a pictorial review of screen-detected cancers and false recalls attributed to tomosynthesis in prospective screening trials. *Breast* 2016; **26**: 119–34.
- Lång K, Andersson I, Rosso A, Tingberg A, Timberg P, Zackrisson S. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmö Breast Tomosynthesis Screening Trial, a population-based study. *Eur Radiol* 2016; **26**: 184–90.
- Houssami N, Skaane P. Overview of the evidence on digital breast tomosynthesis in breast cancer detection. *Breast* 2013; **22**: 101–08.
- Perry N, Broeders M, de Wolf C, Tornberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. *Ann Oncol* 2008; **19**: 614–22.