The Breast 43 (2019) 59–66

Contents lists available at ScienceDirect

The Breast

journal homepage: www.elsevier.com/brst

Original article

Agreement between digital breast tomosynthesis and pathologic tumour size for staging breast cancer, and comparison with standard mammography

M. Luke Marinovich [a, *], Daniela Bernardi [b], Petra Macaskill [a], Anna Ventriglia [b], Vincenzo Sabatino [b], Nehmat Houssami [a]

[a] Sydney School of Public Health, Sydney Medical School, Edward Ford Building (A27), The University of Sydney, NSW 2006, Australia
[b] U.O. Senologia Clinica e Screening Mammografico, Dipartimento di Radiodiagnostica, APSS Trento, Italy

ARTICLE INFO

Article history:
Received 15 June 2018
Received in revised form
29 October 2018
Accepted 2 November 2018
Available online 10 November 2018

Keywords:
Tomosynthesis
Mammography
Breast cancer
Staging
Tumour size
Accuracy

ABSTRACT

Purpose: Tomosynthesis is proposed to improve breast cancer assessment and staging. We compared tomosynthesis and mammography in estimating the size of newly-diagnosed breast cancers.
Methods: All pathologically-confirmed cancers detected in the STORM-2 trial (90 cancers, 85 women) were retrospectively measured on tomosynthesis by two independent readers. One reader also measured cancers on mammography. Relative mean differences (MDs) and 95% limits of agreement (LOA) with pathology were estimated for tomosynthesis and mammography within a single reader (Analysis 1) and between two readers (Analysis 2).
Results: Where cancers were detected and hence measured by both tests, tomosynthesis overestimated pathologic size relative to mammography (Analysis 1: MD 5% versus 1%, Analysis 2: 7% versus 3%; $P = 0.10$ both analyses). There was similar, large measurement variability for both tests (LOA range: −60% to +166%). Overestimation by tomosynthesis was attributable to the subgroup with dense breasts (MDs = 12–13% versus 4% for mammography). There was low average bias for both tests in the low-density subgroup (MDs = 0–4%). LOA were larger in dense breasts for both tomosynthesis and mammography ($P \leq 0.02$ all comparisons). Cancers detected only by tomosynthesis were more frequently in dense breasts (60–68%): for those tumours size was estimated with increased measurement variability (LOA ranging from −75% to +293%).
Conclusions: On average, tomosynthesis overestimates pathologic tumour size in women with dense breasts; that difference is more likely to impact management in women with larger tumours. The main advantage of tomosynthesis appears to be detecting mammographically-occult cancers; however tomosynthesis less accurately measured those cancers in dense breasts (large measurement variability).

1. Introduction

Digital breast tomosynthesis is a recent evolution of full-field digital two-dimensional (2D) mammography involving three-dimensional (3D) transformation of breast images. By acquiring images in a series of thin slices through the breast, tomosynthesis (3D-mammography) has the potential to improve visualisation of breast cancer (BC) due to a reduction in the effect of tissue superimposition on standard 2D-mammography [1]. Current tomosynthesis technology allows for software reconstruction of synthetic 2D-mammography images from 3D acquisitions, thereby roughly halving the radiation dose necessary to obtain separate 2D and 3D scans [2].

Tomosynthesis in the screening setting has been evaluated in prospective studies that show detection of additional cancers when the test is added to mammography [3–5]. Given such improvements in BC detection, the application of tomosynthesis in pre-treatment assessment of cancer extent is an important area for further study, particularly given the increasing adoption of tomosynthesis into population screening practice. Current guidelines for the diagnosis and evaluation of BC note that tomosynthesis may improve accuracy in this setting, particularly for women with dense

* Corresponding author.
E-mail address: luke.marinovich@sydney.edu.au (M.L. Marinovich).

https://doi.org/10.1016/j.breast.2018.11.001
0960-9776/

breasts [6], but do not yet recommend its routine use [7]. More accurate measurement of tumour size at initial staging has the potential to better inform surgical management.

A small number of studies have evaluated the accuracy of tomosynthesis in measuring primary BC size in the staging setting, with pathologic tumour size as the reference standard. Compared with 2D mammography, tomosynthesis has been found to show slightly better agreement with pathologic tumour size, but studies are inconsistent in showing a tendency for underestimation [8] or overestimation [9] of tumour size by both tests. In addition, those studies used prototype tomosynthesis technologies, and thus have uncertain applicability to commercially-available units used in current practice. In a study using recent technology, an identical proportion (91%) of 2D-mammography and tomosynthesis measurements agreed with pathologic measurements within ±10 mm [10]; however, measurement errors within that relatively large range may have implications for treatment decisions, and differences between tests may be evident with a more precise margin of error [11]. Studies have reported higher Pearson correlations with pathology for tomosynthesis compared with 2D-mammography [12], but the limitations of such correlations are well-documented: Pearson's correlation does not assess agreement between measurements and may result in misleading conclusions [13].

Given the paucity of previous studies and their above-stated limitations [14], we undertook a retrospective study of tumour size estimation by tomosynthesis, using prospectively collected data from the Screening with Tomosynthesis Or standard Mammography-2 (STORM-2) population-based BC screening trial [5], and applying the recommended analytic methods to assess agreement between imaging and pathologic measurements [13]. Agreement for tomosynthesis is compared with 2D-mammography, both when tests are interpreted by the same reader, or where different readers perform tumour size measurements.

2. Methods

2.1. Subjects

All pathologically confirmed cancers detected in STORM-2 [5], a prospective population-based screening study of BC detection and false positive recall, were included in this analysis. STORM-2 screened 9672 women using two parallel arms whereby double-readings of the same screening examinations were conducted: 1) sequential standard 2D followed by integrated 2D/3D-mammography *versus* 2) sequential synthetic 2D followed by integrated synthetic 2D/3D-mammography. An outline of the STORM-2 trial pathway is presented in Fig. S1 (online Appendix). Participants were asymptomatic women undergoing biennial screening in Trento, Italy. Women were recalled for further assessment based on recall by any reader in either arm. A total of 90 cancers in 85 women were detected and are the subject of this retrospective sub-study of STORM-2 that focuses on tumour size.

2.2. Imaging details

Participants in STORM-2 had digital mammography with 2D and 3D (tomosynthesis) mammography (acquired with Selenia® Dimensions Unit operated in COMBO© mode; Hologic, Bedford MA, USA). Both 2D and 3D images were acquired at the same screening examination with a single breast positioning per view. Mediolateral oblique and cranio-caudal views were obtained for 2D and 3D acquisitions. Additional trial details have been described by Bernardi et al. [5].

Breast density based on 2D-mammographic findings was assessed as the majority score from screen-readers of STORM-2 (or by arbitration, as required) [5] according to the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) classification: BI-RADS 1 (almost entirely fatty); BI-RADS 2 (scattered areas of fibroglandular density); BI-RADS 3 (heterogeneously dense); and BI-RADS 4 (extremely dense). For analysis, those categories were collapsed into “low density” (BI-RADS 1 or 2) and “high density” (BI-RADS 3 or 4) groupings.

2.3. Imaging tumour size evaluation

Two radiologists (AV and VS) with two and three years experience with 2D-mammography and six and 12 months training in 3D-mammography, respectively, participated in this sub-study of STORM-2. Neither radiologist had participated in the screen-readings for STORM-2. One radiologist (Reader A) retrospectively measured tumour size sequentially on 2D (N = 61) and then 3D images (N = 82) for the cancers originally detected by those modalities in arm 1 of STORM-2. Measurements for both tests were undertaken in the same session. The second radiologist (Reader B) read and measured tumour size 3D images (N = 85), for the cancers originally detected by those modalities in arm 2 of STORM-2. Each radiologist was blinded to other imaging results, pathological findings, and the results of the other reader. The maximum tumour diameter was measured visually with a ruler to the nearest millimetre. 3D-mammography measurements were based on the slice with the largest tumour extension. For lesions with long spiculations, the core of the lesion was measured. For lesions depicted as microcalcifications, the maximum extent was measured.

2.4. Reference standard

All cancers were confirmed at excision histopathology, with reporting by dedicated breast pathologists. Tumour size was based on the maximum invasive tumour diameter, or the extent of DCIS for DCIS only cases, as measured on macroscopic examination by the pathologist and described in the final histopathology report.

2.5. Statistical analysis

Patient characteristics were summarised descriptively with means for age and percentages for categorical variables (mammographic features; breast density; ER/PR status; HER2 status).

Two parallel analyses of tumour size were undertaken. In *Analysis 1*, measurements from mammography and tomosynthesis by Reader A were compared with pathology (within single reader analysis). In *Analysis 2*, Reader A's measurements from mammography and Reader B's measurements from tomosynthesis were compared with pathologic measurements (between-readers analysis). For both analyses, Bland-Altman scatterplots of the absolute differences and the differences of log-transformed measurements between imaging and pathology (vertical axis) and their mean (horizontal axis) were constructed for tumours that were detected (and hence could be measured) by both tomosynthesis and mammography. Plots were examined to assess whether the differences were normally distributed [13]. Simple linear regression was used to investigate association between the underlying size of the measurements and the variation of the residuals around the line of best agreement [15]. Measurement biases for mammography and tomosynthesis were estimated as the *absolute* mean differences (MDs) between those tests and pathology, with positive MDs indicating an average systematic bias towards overestimation of pathologic size, and negative MDs indicating an average bias towards underestimation [13]. Associated 95% limits of agreement (LOA) were also calculated. *Relative* MDs were derived by exponentiation of the mean difference of log-transformed

measurements. Paired sample t-tests were used to compare MDs for each test.

When tumours were detected and measured by only one of the tests (e.g. tomosynthesis but not mammography), descriptive characteristics for those women were compared to the group with cancers detected by both tests using independent sample t-tests for age as a continuous variable, and chi-squared or Fisher's exact tests for categorical variables. MD and LOA with pathology were also calculated separately for this group, and compared to the group with cancers detected by both tests using independent sample t-tests (MD), and the F statistic to test for equality of variances (LOA) [16].

Percentages of agreement between tests and pathology (and associated percentages of over- and underestimation) were also calculated, using ±0 mm, 1 mm, 2 mm, 5 mm, and 10 mm as "margins of error" to define agreement. Proportions of agreement (test measurement within the "margin of error") versus disagreement (test measurement not within the "margin of error", including both over- and underestimation) for were compared with McNemar's test.

All tests of statistical significance were two-sided. The level chosen for statistical significance was $p < 0.05$; $p < 0.10$ was considered to represent weak evidence of a difference. Analyses were undertaken in SAS version 9.4.

3. Results

Eighty-five women including five with bilateral cancers (90 cancers total) were included in this sub-study of STORM2. Characteristics of the cohort are presented in Table 1. The mean age was 59.3 years (median 61.0, range 49–69); 41% of cancers were detected in women who had dense breasts. All cancers were detected on tomosynthesis (3D) images in at least one of the reading arms of STORM-2 (82 [91%] in arm 1 and 85 (94%) in arm 2); 77 cancers [86%] were detected on tomosynthesis in both arms. There were 61 cancers (68%) identified on 2D-mammography in STORM-2; tomosynthesis detected all 61 of those cancers in arm 1, and 58 in arm 2. Cancer detection data have been detailed in our report of the STORM-2 trial [5].

Bland-Altman plots of the absolute differences between the tests and pathology against their mean are presented in Fig. S2 (online Appendix). In both Analyses 1 and 2, and for both mammography and tomosynthesis, the plots suggested a tendency for larger differences between imaging and pathology measurements with increasing tumour size. There was no evidence of association between the differences and their mean ($p \geq 0.40$ for all tests and Analyses); however, there was strong evidence for greater variability in measurement with increasing tumour size ($p \leq 0.0001$ all analyses). Log transformation of the differences resulted in no association with underlying tumour size for either the difference in measurements ($p \geq 0.30$ for all) or their variability ($p \geq 0.40$ for all) (see Fig. 1). Hence, analyses of relative (log) differences are reported here; analyses of absolute differences are presented in Table S1 (online Appendix) to allow comparison with previous studies presenting those data.

Table 1
Characteristics of breast cancers.

Patient characteristics	N (breasts)	Mean (range) or percent
Age[a]		59.3 (49–69)[a]
Mammographic features		
Distortion	15	17%
Microcalcification	19	21%
Irregular margins	34	38%
Stellate	22	24%
Density		
Low	53	59%
High	37	41%
ER/PR		
Positive	71	79%
Negative	4	4%
Unknown	15	17%
HER2		
Positive	5	6%
Negative	67	74%
Unknown	18	20%

[a] Based on N = 85 patients with cancer detected. Mean age differs from that reported in STORM-2, which was calculated on the entire screening cohort.

3.1. Cancers identified and measured by both tomosynthesis and mammography

3.1.1. Overall accuracy

Percentages of cases in which the tests agreed with pathology within a specified margin of error (and corresponding percentages of over- and underestimation) are presented in Table 2. As expected, the percentage agreement increased with larger margins of error, from approximately 15% for exact agreement to approximately 90% for agreement within ±10 mm. There were no statistically significant differences in percentage agreement between mammography and tomosynthesis (all $P > 0.18$).

Table 3 presents relative MDs and LOA between tests and pathology for patients with tumour measurements by both tomosynthesis and mammography; comparisons are presented for within single reader measurements (mammography and tomosynthesis interpreted by the same reader: Analysis 1) and between-readers measurements (tests interpreted by different readers: Analysis 2). Systematic bias in measuring tumour size was relatively small for both tests, with positive values for the MD indicating, on average, overestimation of pathologic tumour size (MD with pathology of 1% for mammography versus 5% for tomosynthesis in Analysis 1; 3% versus 7% in Analysis 2). There was weak statistical evidence that overestimation relative to pathology was greater for tomosynthesis than for mammography ($P = 0.10$ for both Analyses). LOA with pathologic measurements were comparable for mammography and tomosynthesis, and indicated relatively large variability in measurement for both tests (Table 3).

3.2. Stratification by breast density

Stratification by breast density indicated that differences in systematic bias between tomosynthesis and mammography were attributable to the subgroup with dense breasts (Table 3). In both analyses, and relative to mammographic measurements (MD of 4% both Analyses), there was evidence of greater overestimation of pathologic tumour size by tomosynthesis in the high-density cohort (MD of 13% in Analysis 1 ($P = 0.03$); MD of 12% in Analysis 2 ($P = 0.07$)). In the low-density cohort, there was no evidence that MDs with pathology were different for mammography (0% in Analysis 1; 2% in Analysis 2) and tomosynthesis (2% in Analysis 1; 4% in Analysis 2: $P > 0.59$ for difference between tests in both Analyses). MDs indicated relatively low systematic measurement bias in this cohort. Within density strata, LOA were comparable between tests (Table 3).

There was no evidence that MDs differed between low- and high-density strata for either tomosynthesis ($P \geq 0.49$ for both Analyses) or mammography ($P \geq 0.82$); however, LOA were larger in the high-density cohort versus the low-density cohort (Table 3), with statistical evidence in both Analyses for greater variance in measurement in dense breasts for both tomosynthesis and mammography ($P \leq 0.02$ for all comparisons).

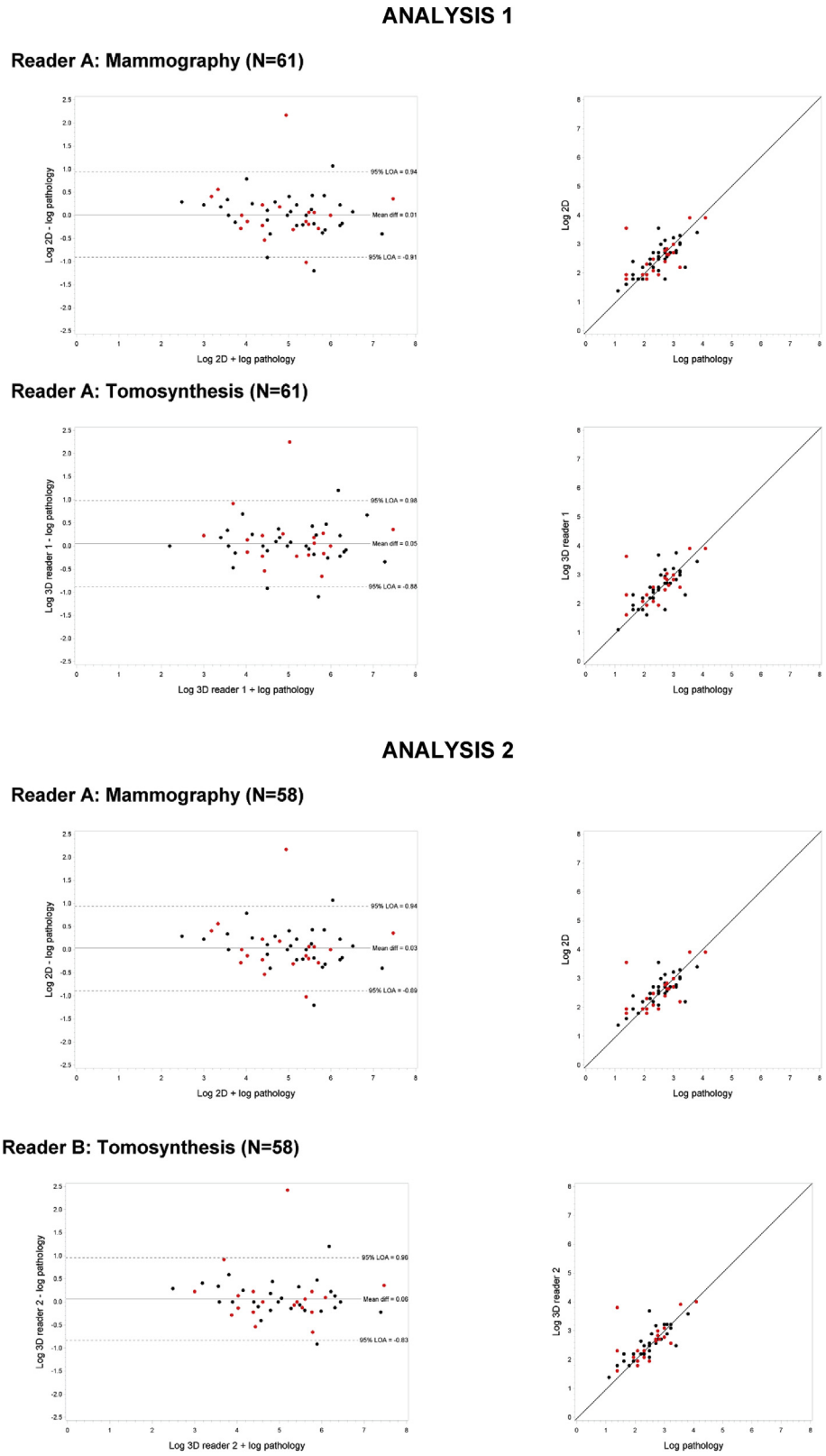

Fig. 1. Bland-Altman plots for *relative* MDs and LOA for test and pathology (tumours detected by both tests), for Analysis 1 (within single reader) and Analysis 2 (between readers). Black dots represent women with low breast density; red dots represent women with high breast density. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Percentage agreement (within specified margin of error) between tumour measurement by tests and pathology: Tumours measured by both tests.

Margin of error	Agreement % (Underestimation, Overestimation)			
	Analysis 1[a] (N = 61)		Analysis 2[a] (N = 58)	
	Mamm (Reader A)	Tomo (Reader A)	Mamm (Reader A)	Tomo (Reader B)
0 mm	13 (44,43)	13 (44,43)	14 (43,43)	17 (45,38)
1 mm	33 (36,31)	41 (30,30)	31 (36,33)	41 (29,29)
2 mm	54 (28,18)	51 (26,23)	53 (28,19)	62 (19,19)
5 mm	79 (11,10)	80 (8,11)	79 (10,10)	86 (5,9)
10 mm	90 (5,5)	89 (5,7)	90 (5,5)	91 (3,5)

Note: Percentages may not sum to 100% due to rounding.
[a] Analysis 1 compares measurements within a single reader; Analysis 2 compares measurements between two readers.

Table 3
Relative (%) mean difference (MD) with pathology and limits of agreement (LOA) for 2D mammography and tomosynthesis.

	Whole study cohort			Low breast density			High breast density		
	N	MD (%)	LOA (%)	N	MD (%)	LOA (%)	N	MD (%)	LOA (%)
Analysis 1: Within single reader measurements									
Cancers measured by both tests									
2D mammography (Reader A)	61	1[a]	−60, +155	41	0	−54, +116	20	4[b]	−69, +244
3D tomosynthesis (Reader A)	61	5[a]	−58, +166	41	2	−53, +122	20	13[b]	−66, +271
Cancers measured by one test only									
2D mammography (Reader A)	0	–	–	0	–	–	0	–	–
3D tomosynthesis (Reader A)	21	1	−72, +260	7	8	−65, +237	14	−3	−75, +283
Analysis 2: Between reader measurements									
Cancers measured by both tests									
2D mammography (Reader A)	58	3[a]	−59, +156	38	2	−51, +114	20	4[c]	−69, +244
3D tomosynthesis (Reader B)	58	7[a]	−57, +161	38	4	−46, +100	20	12[c]	−68, +290
Cancers measured by one test only									
2D mammography (Reader A)	3	−26	−75, +124	3	−26	−75, +124	0	–	–
3D tomosynthesis (Reader B)	27	3	−69, +240	11	0	−65, +185	16	5	−72, +293

[a] p = 0.10.
[b] p = 0.03.
[c] p = 0.07.

Table 4 presents relative MDs and LOA applied to the minimum (3 mm), median (12 mm) and maximum (60 mm) values of pathologic size observed in our data, to derive the mean imaging size and LOA in mm for each value. For tumours measured by both tomosynthesis and mammography, the mean imaging sizes indicate that systematic bias for both tests is clinically comparable, with greater tendency for overestimation of pathologic size by tomosynthesis only for larger tumours in women with dense breasts. However, for both tests, LOA indicate a risk of 'overstaging' for medium sized tumours (the upper range of the LOA crossing the threshold of 20 mm, e.g. upstaging from Stage I to Stage II) based on imaging size, and 'understaging' for larger tumours. The latter was especially pronounced for women with high breast density, in whom the lower extreme of the LOA indicated possible 'downstaging' from Stage III (≥50 mm) to Stage I (<20 mm).

3.3. Sensitivity analysis

In sensitivity analyses excluding two cases with extensive DCIS components at pathological examination (one each in the low- and high-density subgroups), there was minimal mean bias observed for tomosynthesis (MD = −1% in Analysis 1; 0% in Analysis 2). There was larger mean underestimation by mammography (−4% in Analysis 1; −3% in Analysis 2), but the difference with tomosynthesis was not statistically significant ($P \geq 0.14$) (Table S2, online Appendix). For the high-density subgroup, there was statistical evidence for greater mean underestimation by mammography (MD = −7%) compared with tomosynthesis (1%) in Analysis 1 ($P \leq 0.04$). A similar finding in Analysis 2 (−7% versus 1%) did not reach statistical significance ($P = 0.11$), but the small number of patients in this subgroup (N = 19) limits power to detect a difference between tests.

Relative to the main analysis, smaller LOA were observed that were comparable for mammography and tomosynthesis. LOA were also comparable across density strata, with no evidence of inequality in variances between low- and high-density subgroups for either tomosynthesis ($P \geq 0.29$) or mammography ($P \geq 0.73$).

3.4. Cancers identified and measured by tomosynthesis only

In both Analyses 1 and 2, a proportion of cancers were measured by tomosynthesis only, comprising cancers not detected at mammography (21 cancers [23%] in Analysis 1; 27 cancers [30%] in Analysis 2) (Table 3). Compared with cancers measured by both tests, the additional cases identified and measured by tomosynthesis were younger (Analysis 1: mean 56 versus 61 years, $P = 0.003$; Analysis 2: mean 55 versus 61 years, $P = 0.0005$) and were more likely to have high breast density (Analysis 1: 67% vs 33%, $P = 0.007$; Analysis 2: 59% vs 34%, $P = 0.03$). In contrast, few cancers were measured by mammography only (N = 3 [3%] in Analysis 2, all in low-density breasts).

For the additional cases measured by tomosynthesis but not mammography, there was only small overestimation of tumour size (Analysis 1: MD 1%; Analysis 2: 3%); MDs were similar to those from cases measured by both tests (both $P \geq 0.73$). However, compared with tumours measured by both tests, there was greater measurement variability in cases measured by tomosynthesis but not mammography (Analysis 1: LOA -72%, +260%; Analysis 2: LOA -69%, +240%), with statistical evidence of inequality in variances (both $P \leq 0.07$) (Table 3).

Table 4
Comparison of differences between test and pathology for minimum (3 mm), median (12 mm) and maximum (60 mm) pathologic tumour sizes in the data set.

	Pathologic tumour size								
	3 mm			12 mm			60 mm		
	Whole cohort	Low density	High density	Whole cohort	Low density	High density	Whole cohort	Low density	High density
Analysis 1: Within single reader measurements									
Tumours detected by both tests									
2D (Reader A)									
Mean 2D size (mm)	3	3	3	12	12	12	61	60	62
LOA: low, high (mm)	1,8	1,6	1,10	5,31	6,26	4,42	24,153	28,130	19,206
LOA: range (mm)	7	5	9	26	20	38	129	102	187
3D (Reader A)									
Mean 3D size (mm)	3	3	3	13	12	14	63	61	68
LOA: low, high (mm)	1,8	1,7	1,12	6,33	6,27	6,46	25,160	28,133	20,223
LOA: range (mm)	7	6	11	27	21	40	135	105	203
Tumours detected by 3D only									
3D (Reader A)									
Mean 3D size (mm)	3	3	3	12	13	12	61	65	58
LOA: low, high (mm)	1,11	1,10	1,11	3,43	5,41	3,46	17,217	26,207	13,228
LOA: range (mm)	10	9	10	40	36	43	200	181	215
Analysis 2: Between reader measurements									
Tumours detected by both tests									
2D (Reader A)									
Mean 2D size (mm)	3	3	3	12	12	12	62	61	62
LOA: low, high (mm)	1,8	2,6	1,10	5,31	6,26	4,42	25,154	29,128	19,206
LOA: range (mm)	7	4	9	26	20	38	129	99	187
3D (Reader B)									
Mean 3D size (mm)	3	3	3	13	12	13	64	62	67
LOA: low, high (mm)	2,8	2,6	1,12	6,32	7,24	5,48	26,157	32,120	19,234
LOA: range (mm)	6	4	11	26	17	43	131	88	215
Tumours detected by 3D only									
3D (Reader A)									
Mean 3D size (mm)	3	3	3	12	12	13	62	60	63
LOA: low, high (mm)	1,10	1,9	1,12	4,41	4,34	4,48	20,206	21,171	20,239
LOA: range (mm)	9	8	11	37	30	44	186	150	219

4. Discussion

In the assessment of newly diagnosed BC, accurate measurement of the size of the primary tumour assists in staging and guiding surgical management. We undertook a retrospective substudy of cancers detected in the STORM-2 screening trial [5], and found that tomosynthesis showed a small overall bias towards overestimation of tumour size (MD = 5–7%) relative to mammography (MD = 1–3%) when cancers were detected by both tests. In stratified analyses based on breast density, the observed overestimation by tomosynthesis was attributable to the high-density subgroup; for those women, MDs (12–13%) were statistically significantly greater than for mammography (4%). Clinicians should be aware of the potential for tomosynthesis to overestimate pathologic tumour size in women with dense breasts, particularly for those with larger tumours where proportional bias translates to greater absolute overestimation (Table 4).

It is noteworthy that sensitivity analysis suggested that such overestimation may be due to extensive DCIS components visualised and measured on tomosynthesis but not included in pathologic measurements of the index cancer. This is consistent with previous findings that inclusion or exclusion of DCIS components in pathologic tumour measurements may affect agreement with imaging tests [11]. Exclusion of those few cases resulted in comparable MDs between tests in the low density stratum, but with statistical evidence of underestimation by mammography (MD = −7%) relative to tomosynthesis (MD = 1%) in dense breasts ($P < 0.04$ in Analysis 1). Since breast cancer T-staging is based on the measurement of invasive cancer only, our data highlight the challenge of distinguishing the extent of invasive tumour from DCIS components using mammography and also using tomosynthesis.

Variability in measurement was comparable for both tomosynthesis and mammography, but was greater in dense breasts for both tests. For women with high breast density, 95% LOA indicated that the "true" (pathologic) tumour size may be up to 166% above or 60% below the imaging measurement. Sensitivity analysis again suggested that this relatively large range of measurement error was mostly attributable to a small number of cases with extensive DCIS components reported on histology. When two cases with extensive DCIS at pathology were excluded (one low and one high density), LOA for both tomosynthesis and mammography decreased markedly in women with dense breasts, and were comparable to LOA in the low density subgroup (which were also slightly lower compared with the main analysis). However, even within this more limited range, measurement errors may potentially impact management when treatment decisions are informed by these tests. For example, underestimation may result in inadequate resection requiring reintervention for positive margins; overestimation may lead to upstaging and/or a change in treatment approach (e.g. mastectomy instead of breast conserving surgery).

A proposed advantage of tomosynthesis is the reduction of tissue superimposition apparent on mammographic images, thereby potentially improving imaging of malignancy in women with dense breasts. Studies in the screening setting provide evidence for this advantage, with tomosynthesis demonstrated to detect additional cancers (not detected at mammography) particularly in high-density subgroups [6]. However, our study suggests that the enhanced detection capability of tomosynthesis does not

necessarily translate into enhanced measurement of tumour size. For cancers detected only by tomosynthesis, frequently in women with dense breasts, tumour size measurement had substantial measurement variability, with LOA that were larger than corresponding estimates for cancers measured by both tests. This is likely to reflect the fact that mammographically-occult cancers represent "challenging" cases, with tumour and breast density characteristics that make detection and imaging size measurement inherently difficult. Based on these results, tomosynthesis does not appear to improve size measurement when standard mammographic images also visualise the cancer; for cancers detected by tomosynthesis but not mammography, there is potential that those cancers may be mis-sized.

Previous studies have used statistical methods that may lead to misleading conclusions about the accuracy of tomosynthesis for measuring tumour size, or have not used commercially available tomosynthesis technology, with unknown applicability to current practice. Our study addresses all those limitations, and further extends the evidence base by reporting relative mean differences, which provide better estimates of measurement error with increasing tumour size; stratifying analyses by density; exploring possible causes of measurement inaccuracy through sensitivity analysis; comparing tomosynthesis and mammography using paired measurements; and assessing measurement accuracy for cancers detected by tomosynthesis alone. However, a limitation of our study was that radiologists read mammography and tomosynthesis images sequentially, meaning that interpretation of tomosynthesis was not blinded to mammography. This may have resulted in tomosynthesis size measurements being more similar to mammography than if readers were blinded to mammographic measurement. There is therefore the potential for our study to have underestimated the differences between tests.

The clinical application and evaluation of tomosynthesis in breast cancer imaging and diagnosis has emerged rapidly in recent years, and includes its application in screening, assessment, and staging [17–19]. To date, there are no guidelines recommending its use in place of mammography in breast screening or diagnosis, however there is strong evidence that using tomosynthesis in population breast screening improves detection metrics relative to mammography screening. A comprehensive meta-analysis [17] has estimated the incremental cancer detection rate from tomosynthesis to be 1.6 cancers per 1000 screens, with an absolute reduction in recall of 2.2%, compared to 2D mammography. In addition, there are several RCTs of tomosynthesis screening in progress [20]. There is relatively less evidence on use of tomosynthesis in diagnosis, as highlighted in a recent review which reported that tomosynthesis for assessment or investigation of screen-detected abnormalities and lesions had been evaluated in few studies (with methodological limitations) and these studies generally show that tomosynthesis can potentially improve specificity over mammography [21].

Our findings indicate that tomosynthesis – at least for the current version of the technology – does not improve the measurement of BC size relative to standard mammography, when cancers are seen on both tests. Tomosynthesis detects additional, mammographically-occult lesions in the screening setting, but measurement variability is large, and hence clinicians should be aware of the possibility of inaccurate measurement by tomosynthesis in those women. Given that the number of cancers detected only by tomosynthesis was modest, our finding of relatively inaccurate measurement in that group requires further exploration in larger studies, and should be interpreted in the context that tomosynthesis-only detection of BC was frequently in dense breasts which also limits cancer size measurement.

Funding

This work was supported by a Cancer Institute NSW (CINSW) Early Career Fellowship, Grant ID No. 14/ECF/1-06 (ML Marinovich), and by a National Breast Cancer Foundation (NBCF Australia) Breast Cancer Research Leadership Fellowship (N Houssami).

The study sponsors had no role in the study design, collection, analysis and interpretation of data; in the writing of the manuscript; and in the decision to submit the manuscript for publication.

Conflicts of interest

All authors declare no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.breast.2018.11.001.

References

[1] Houssami N, Skaane P. Overview of the evidence on digital breast tomosynthesis in breast cancer detection. [Review]. Breast 2013;22:101–8.
[2] 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. [Review]. Breast 2015;24:93–9.
[3] 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.
[4] 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.
[5] Bernardi D, Macaskill P, Pellegrini M, et al. Breast cancer screening with tomosynthesis (3D mammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study. Lancet Oncol 2016;2016:2016. Date of Publication: 2016.
[6] Houssami N, Turner RM. Rapid review: estimates of incremental breast cancer detection from tomosynthesis (3D-mammography) screening in women with dense breasts. Breast 2016;30:01.
[7] Senkus E, Kyriakides S, Ohno S, et al. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2015;26:mdv298.
[8] Fornvik D, Zackrisson S, Ljungberg O, et al. Breast tomosynthesis: accuracy of tumor measurement compared with digital mammography and ultrasonography. Acta Radiol 2010;51:240–7.
[9] Seo N, Kim HH, Shin HJ, et al. Digital breast tomosynthesis versus full-field digital mammography: comparison of the accuracy of lesion measurement and characterization using specimens. Acta Radiol 2014;55:661–7.
[10] Destounis SV, Arieno AL, Morgan RC. Preliminary clinical experience with digital breast tomosynthesis in the visualization of breast microcalcifications. J Clin Imag Sci 2013;3:65.
[11] Marinovich ML, Macaskill P, Irwig L, et al. Meta-analysis of agreement between MRI and pathologic breast tumour size after neoadjuvant chemotherapy. Br J Canc 2013;109:17.
[12] Meacock LM, Mombelloni AI, Iqbal A, Akbar N, Wang Y, Michell MJ. The accuracy of breast cancer size measurement: digital breast tomosynthesis (DBT) vs 2D digital mammography (DM). Insights Imagin 2010;1:S306.
[13] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:1986.
[14] Marinovich ML, Macaskill P, Bernardi D, et al. Systematic review of agreement between tomosynthesis and pathologic tumor size for newly diagnosed breast cancer and comparison with other imaging tests. Expet Rev Med Dev 2018;15(7):489–96.
[15] Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:1999.
[16] Woodward M. Basic analytic procedures. Epidemiology: study design and data analysis. London: Chapman & Hall/CRC; 1999. p. 31–105.
[17] Marinovich ML, Hunter KE, Macaskill P, et al. Breast cancer screening using tomosynthesis or mammography: a meta-analysis of cancer detection and recall. J Natl Cancer Inst 2018;110(9):942–9.
[18] Gilbert FJ, Tucker L, Young KC. Digital breast tomosynthesis (DBT): a review of the evidence for use as a screening tool. Clin Radiol 2016;71(2):141–50.

[19] Thibault F, Dromain C, Breucq C, et al. Digital breast tomosynthesis versus mammography and breast ultrasound: a multireader performance study. Eur Radiol 2013;23(9):2441–9.
[20] Aase HS, Holen AS, Pedersen K, et al. 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 trial. Eur Radiol 2018. https://doi.org/10.1007/s00330-018-5690-x.
[21] Li T, Marinovich ML, Houssami N. Digital breast tomosynthesis (3D mammography) for breast cancer screening and for assessment of screen-recalled findings: review of the evidence. Expet Rev Anticancer Ther 2018;18(8):785–91.