Research article

Effect of integrating digital breast tomosynthesis (3D-mammography) with acquired or synthetic 2D-mammography on radiologists' true-positive and false-positive detection in a population screening trial: A descriptive study

Daniela Bernardi[a], Tong Li[b,*], Marco Pellegrini[a], Petra Macaskill[b], Marvi Valentini[a], Carmine Fantò[a], Livio Ostillio[a], Nehmat Houssami[b,*]

[a] U.O. Senologia Clinica e Screening Mammografico, Department of Diagnostics, Ospedale di Trento, Azienda Provinciale Servizi Sanitari, Trento, Italy
[b] Sydney School of Public Health, Faculty of Medicine and Health, University of Sydney, Sydney, Australia

ABSTRACT

Background: We previously reported the Screening with tomosynthesis or standard mammography-2 (STORM-2) trial, showing that tomosynthesis (3D-mammography) screening detected more cancers than 2D-mammography in double-reading practice. In this study, we report reader-specific detection measures for radiologists who performed the screen-reading in this trial.

Methods: This is a sub-study of the STORM-2 trial which prospectively integrated 3D-mammography with acquired or synthetized 2D-mammograms in parallel double-reading arms. Asymptomatic women ≥ 49 years who attended population-based screening (Trento, 2013–2015) were recruited. Screening participants were recalled at any positive sequential screen-read in either reading arm of the trial. Radiologist-specific detection measures were calculated for each of seven radiologists who performed screen-reads: number of detected cancers, proportion of true-positive (TP) detection, and number and rate of false-positive (FP) recalls (FPR). We estimated incremental cancer detection rate (CDR) from integrating 3D-mammography in screen-reading.

Results: Across all radiologists, TP detection (relative sensitivity) ranged between: 46% and 100% (median 59.5%) for 2D-mammography; 75% and 100% (median 76%) for integrated 2D/3D-mammography screening; 56% and 76% (median 64%) for 2Dsynthetic; 67% and 88% (median 78%) for 2Dsynthetic/3D-mammography. Integrating 3D-mammography led to incremental CDRs between 0/1000 and 3.5/1000 screens. FPR ranged between: 1.2% and 2.7% (median 2.25%) for 2D-mammography; 1.5% and 3.4% (median 2.75%) for 2D/3D-mammography; 1.6% and 4.6% (median 2.4%) for 2Dsynthetic; and 1.8% and 6.7% (median 3.0%) for 2Dsynthetic/3D-mammography.

Conclusions: There was variability in the magnitude of effect from integrating 3D-mammography (relative to screen-reading with acquired or synthesised 2D-mammography alone) on individual radiologist's TP and FP detection, although there was an overall pattern of increasing cancer detection and also increasing FP recall for most readers.

1. Background

Mammography technology has evolved through the development of digital breast tomosynthesis, a pseudo-three-dimensional mammography technology (also referred to as 3D-mammography) which appears to be increasingly adopted in screening practice. Tomosynthesis has been evaluated in prospective trials [1–4] within European population-based screening programs, and in retrospective studies

conducted in North America [5–10], all of which demonstrate that it enhances detection measures although results are heterogeneous across screening settings. Randomised controlled trials of this new mammography technology have also been initiated. We recently reported results of the screening with tomosynthesis or standard mammography-2 (STORM-2) population-based trial in which breast tomosynthesis (3D-mammography) was interpreted either with acquired 2D images or with reconstructed images (synthesised 2D) – our trial showed that screening

using 3D detected more breast cancers than 2D-mammography based on standard double-reading practice [4].

In the present study, we focus on reader-specific interpretation in STORM-2 to elucidate the effect of integrating tomosynthesis in screening practice on individual radiologists' detection capability. Prospective studies of tomosynthesis screening have reported little information on reader-specific data (given they are based on double-reading) or have reported interim analyses [2–4,11]. Therefore, in the present analysis, we aim to provide insights into individual screen-readers' performance with the addition of tomosynthesis alongside either acquired or synthesised 2D-mammograms, and to identify potential variability in the effect of integrating tomosynthesis in screening.

2. Methods

This work is a secondary analysis of STORM-2, a prospective population-based screening trial that compared mammography screen-reading in *sequential* phases in two parallel double-reading arms as described in our primary report: screen-reading in one double-reading arm used acquired 2D and 3D mammography, and another double-reading arm used 3D-mammography with synthesised 2D-images (2Dsynthetic). The trial methodology, reported by Bernardi et al. [4], provides paired data for each screening examination (from the same participant) in each arm of the trial. Asymptomatic women aged ≥ 49 years attending biennial population-based screening through the Trento screening program, Italy, were recruited into the study between May 2013 and May 2015, and invited to have screening with 2D and 3D mammography. The study was granted institutional ethics approval, and consent was obtained from screening participants [4].

2.1. Mammography and screen-readings

Participants had digital mammography integrating both 2D and 3D mammography acquisitions using Selenia® Dimensions Unit operated in COMBO© mode; Hologic, Bedford MA, USA and using C-View™ 2D-software to reconstruct 2D-mammographic images from 3D acquisitions. All mammography acquisitions were obtained 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. Women who declined to participate in the trial had 2D-mammography. As reported in our earlier publication, the estimated mean glandular dose *per view* was 1.36 mGy (SD 0.51) from 2D-mammography acquisition, 1.87 mGy (SD 0.67) from 3D-mammography, and 3.22 mGy (SD 1.16) from dual-acquisitions (2D + 3D).

In STORM-2, in one double-reading arm, screens were *sequentially* reported by radiologists viewing 2D-mammography alone, and then re-interpreted by the same radiologists (on the same day) using integrated 2D/3D-mammography. In another *independently* reported double-reading arm, the same screening examinations were interpreted *sequentially* by a different reader pair using 2Dsynthetic and re-reported on the same day using integrated 2Dsynthetic/3D-mammography [4]. Hence each screen was interpreted by two different reader pairs (a total of 4 readings) using 2D/3D or 2Dsynthetic/3D. Radiologists reported each screening mammogram independently of each other, and according to the above-described sequence, readers were asked to record whether or not to recall at each screen-reading phase. A screen was considered positive and the woman recalled to assessment (to have further investigations) if recalled by either screen-reader in either of the double-reading arms, based on recall at any screen-reading phase.

Seven breast radiologists participated in screen-reading and had an average 13 (range 3–23) years' experience in breast imaging. They had received training in 3D-mammography and had been using 3D-mammography an average 2.7 (range 2–3) years at trial initiation. All but one of the 7 radiologists participated in both double-reading arms; one radiologist undertook screen-reading in the 2Dsynthetic/3D double-reading only due to screen-reading scheduling that was constrained by

the independent reading pairing [4]. Previous mammograms were displayed, where available, at the time of screen-reading.

2.2. Outcome measures

Outcome measures were (a) the number and percentage of detected cancers at each screen-reading phase, for each reader, and the incremental cancer detection rate (CDR) per 1000 screens attributed to integrating 3D-mammography in screen-reading; and (b) the number and percentage of false-positive (FP) recalls at each screen-reading phase, for each reader, and the trade-off between the additional FP and true-positive (TP) detection (FP:TP) attributed to integrating 3D-mammography in screen-reading [12].

Outcomes were ascertained on the basis of excision histology, or based on all investigations performed at assessment (additional imaging, and histology from core needle biopsy where performed) in recalled subjects, and included two year follow-up to identify interval cancers.

2.3. Statistical analysis

STORM-2 sample size was planned on the basis of the study's primary end-point (comparison of CDR for double-readings) and has been described in our earlier publication [4]. The present study reports a secondary analysis of reader-specific data; because each screen-reader interpreted only a subset of screens, comparisons are not appropriate – instead we report estimates of detection at each reading phase. For each radiologist, and for each screen-reading phase, we calculated the following: the number of interpreted screens, the number and percentage of detected breast cancers from cancers identified in the study participants (screen-detected plus interval cancers), relative sensitivity (for the cases that an individual reader actually read) at each reading phase, incremental CDR for 3D-mammography, number and percentage of FP recall, and the FP:TP detection [12]. We assessed association between incremental CDR and radiologist experience in screen-reading mammography using the Pearson/Spearman correlation coefficient.

Descriptive analyses and correlation analyses were conducted using SPSS (IBM SPSS statistics for windows, version 24.0). Exact (Clopper-Pearson) 95% confidence intervals (CI) were computed for all percentages and rates per 1000 using SAS/STAT 9.4 (SAS Institute Inc., Cary, NC, USA).

3. Results

STORM-2 included 9672 screening participants (median age 58-years with IQR of 53–63 years), in whom 90 breast cancers were screen-detected (in 85 participants) and 6 interval cancers have been identified at follow-up. The trial's CDR and the characteristics of the breast cancers have been reported [4].

3.1. Radiologist-specific cancer detection

Tables 1 and 2 present cancer detection data for each screen-reader in the two reading arms (which used 2D/3D or 2Dsynthetic/3D) in the STORM-2 trial: there was variability in cancer (TP) detection across participating radiologists for each screen-reading phase, in each reading arm. As shown in Table 1, at 2D-mammography screening, TP detection (relative sensitivity) ranged between 46% and 100% (median 59.5%); and at integrated 2D/3D-mammography screening, TP detection (relative sensitivity) was relatively higher, ranging between 75% and 100% (median 76%). For all but one radiologist, screen-reading with 2D/3D-mammography improved breast cancer detection over 2D-mammography alone (Table 1), resulting in *incremental* CDRs that varied between 0/1000 and 3.5/1000 screens (median 1.45/1000). It should be noted that the relative sensitivity is for the cases that an individual reader actually read, and the confidence intervals for all