How Does the Display Luminance Level Affect Detectability of Breast Microcalcifications and Spiculated Lesions in Digital Breast Tomosynthesis (DBT) Images?

Claudio Ferranti, MD, Alessandra Primolevo, MD, Francesco Cartia, MD, Claudia Cavatorta, MSc, Chiara Maura Ciniselli, MSc, Manuela Lualdi, PhD, Silvia Meroni, PhD, Emanuele Pignoli, PhD, Maddalena Plebani, PhD, Claudio Siciliano, Paolo Verderio, PhD, Gianfranco Scaperrotta, MD

Rationale and Objectives: This study evaluates the influence of the calibrated luminance level of medical displays in the detectability of microcalcifications and spiculated lesions in digital breast tomosynthesis images.

Materials and Methods: Four models of medical displays with calibrated maximum and minimum luminance, respectively, ranging from 500 to 1000 cd/m² and from 0.5 to 1.0 cd/m², were investigated. Forty-eight studies were selected by a senior radiologist: 16 with microcalcifications, 16 with spiculated lesions, and 16 without lesions. All images were anonymized and blindly evaluated by one senior and two junior radiologists. For each study, lesion presence or absence and localization statements, interpretative difficulty level, and overall quality were reported. Cohen’s kappa statistic was computed between monitors and within or between radiologists to estimate the reproducibility in correctly identifying lesions; for multireader-multicase analysis, the weighted jackknife alternative free-response receiver operating characteristic statistical tool was applied.

Results: Intraradiologist reproducibility ranged from 0.75 to 1.00. Interreader as well as reader-truth agreement values were >0.80 and higher with the two 1000 cd/m² luminance displays than with the lower luminance displays for each radiologist. Performances in the detectability of breast lesions were significantly greater with the 1000 cd/m² luminance displays when compared to the display with the lowest luminance value (P value <0.001).

Conclusions: Our findings highlight the role of display luminance level on the accuracy of detecting breast lesions.

Key Words: Digital breast tomosynthesis; breast neoplasm; display device; liquid crystal display; luminance; radiographic image interpretation; ROC curve.

© 2017 The Association of University Radiologists. Published by Elsevier Inc. All rights reserved.

INTRODUCTION

The interpretation of mammographic images may be challenging due to the subtle differences in soft tissue densities of normal and pathological structures of the breast, and diagnosis benefits from the detection of microcalcifications and the assessment of masses’ borders. Mammographic interpretations are based on the evaluation of morphologic descriptors well established by the Breast Imaging Reporting And Data System (BI-RADS) lexicon (1). However, they can be applied only if a lesion is perceived on mammographic studies. The radiologist’s expertise may be hampered if mammographic images are of bad quality or if they are improperly visualized. For this reason, great attention must be given to mammographic images’ presentation devices, both film view box and softcopy display devices, whose wrong choice or improper setup can compromise the overall quality of mammographic examination. Regarding mammographic displays, the ACR-AAPM-SIIM Practice Guideline for Determinants of Image Quality in Digital Mammography recommends that monitors used for interpretation be specifically approved for digital mammography use by the Food and Drug Administration (FDA), with a spatial resolution of 5 megapixel (MP) in a 21” panel (2). Regarding the luminance level, a calibrated maximum luminance of at least 400 cd/m² is required and greater than 450 cd/m² is recommended, but...
an increased display luminance level is welcome, being the human eye ability to detect difference in contrast and fine details affected by overall brightness of a scene. As concerns the calibrated minimum luminance level, no value is specified in the mentioned guidelines.

In literature, several observer performance studies are reported comparing the diagnostic accuracy among displays of different spatial resolutions (3–5), but only a few authors have investigated the role of display luminance level (6–8). Furthermore, to our knowledge, the diagnostic performance of medical displays dedicated to mammography with different setup has been poorly investigated. In the paper of Kimpe and Xthona, an increased level of calibrated maximum luminance is proven to increase the detection probability of breast microcalcifications (9). Briefly, increasing the calibrated maximum luminance of a medical display from 500 cd/m$^2$ to 1000 cd/m$^2$ increases the detection probability of microcalcifications by 13% and 20%, as a function of the parameter used in the Weibull psychometric function. The results presented in the mentioned article have been obtained only on the basis of the Barten’s model of the contrast sensitivity curve, and the authors themselves, in the conclusion of their study, suggested the activation of an observational study to support the presented results.

In this work, we investigated the role of the calibrated luminance level in the detectability of breast microcalcifications and spiculated lesions in digital breast tomosynthesis (DBT) images (10) through the diagnostic assessments performed by three radiologists with different levels of mammographic expertise and by using various medical displays, each with a prefixed setup (ie, different levels of calibrated maximum and minimum luminance).

**MATERIALS AND METHODS**

The study reported in this paper is a part of a larger research started at our Institute, designed to identify the most suitable physical parameters of medical displays for a faithful reproduction of medical images, and to quantify the possible bias in medical image interpretation. The Ethical Committee of our Institute approved the protocol of the study and allowed us to perform a retrospective collection of radiological images and to utilize anonymized image data for the study. Overall, monitors of three different vendors were included in our research to investigate five radiological modalities: digital mammography, tomosynthesis, plain radiography (chest and bone), computed tomography, and nuclear medicine. For each modality, a diagnostic question was chosen and altogether more than 300 images with different levels of difficulty of interpretation were selected by senior radiologists and blindly evaluated by participant radiologists with different levels of expertise. In this paper, we report the study design and the results obtained for DBT image interpretation.

**Equipment and Images**

For the present study on DBT images, four medical displays of the same vendor (Barco N.V., Kortrijk, Belgium), which encompass the whole range of monitors with different characteristics and costs approved for mammographic use, were used; the manufacturer’s specifications of the selected displays, labeled in this paper by letters from “A” to “D,” are summarized in Table 1. The set values are those recommended by the manufacturer for optimal use of the monitor.

As reported in Table 1, to isolate the effect of display luminance on the detectability of breast lesions, all other monitor characteristics and settings were identical, with the exception of monitor D that, despite having the same brightness of the monitor C, was included in our study to evaluate a possible role of a white point generated by a color panel instead of by a monochrome one.

The stock faceplates of the liquid crystal display (LCD) monitors were covered so that no textual information or model number was visible to the observers in the study; only monitor D was recognizable, having a color panel and a different screen size. Each display was initially calibrated by the manufacturer according to the Digital Imaging and Communications in Medicine (DICOM) part 14 grayscale standard display function. Monitor quality control was performed at the beginning of the study and at regular intervals following the AAPM TG18 and the International Electrotechnical Commission (IEC) 62563–1:2009–12 recommendations (11,12) to ensure that no deviation in luminance occurred during the study. All displays were connected to picture archiving and communication system (PACS) workstation used for the reporting of mammographic examinations at our Institute (Dell T7616), equipped with Barco MXRT 5450 or Barco MXRT 7500 graphic controller for displays A and B and C and D, respectively.

**TABLE 1. Physical Properties of the Four Medical Displays Used in Our Study**

<table>
<thead>
<tr>
<th>Display Model and Study Code</th>
<th>Panel</th>
<th>Diagonal Size (cm)</th>
<th>Pixel Pitch (mm)</th>
<th>Maximum Luminance (cd/m$^2$)</th>
<th>Minimum Luminance (cd/m$^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (Barco Nio 5MP MDNG-5221)</td>
<td>LCD monochrome</td>
<td>54.1</td>
<td>0.165</td>
<td>500</td>
<td>0.5</td>
</tr>
<tr>
<td>B (Barco Coronis 5MP MDCG-5221)</td>
<td>LCD monochrome</td>
<td>54.1</td>
<td>0.165</td>
<td>600</td>
<td>0.6</td>
</tr>
<tr>
<td>C (Barco Tomosynthesis 5MP MDMG-5221)</td>
<td>LCD monochrome</td>
<td>54.1</td>
<td>0.165</td>
<td>1000</td>
<td>1</td>
</tr>
<tr>
<td>D (Barco Coronis UNITI 12MP MDMC-12133)</td>
<td>LCD color</td>
<td>85.3</td>
<td>0.169</td>
<td>1000</td>
<td>1</td>
</tr>
</tbody>
</table>

LCD, liquid crystal display.
Forty-eight DBT target studies performed at our Institute from March 2012 to December 2015 were selected by a senior radiologist with more than 30 years of mammographic expertise, showing different levels of interpretative difficulty due to the lesions’ features and sizes, background pattern, and mammographic density. The case selection was performed by using display D, which was the one used by the senior radiologist in his clinical practice at the time of the case selection. The case series included both no-lesion studies (16 no suspicious breast lesions) and lesion studies (16 microcalcifications of interest and 16 spiculated lesions), leading to a total of 34 breast lesions, as two patients presented two different microcalcification localizations (reference evaluation).

The selected studies were confirmed as negative, benign, or malignant by retrieving from the institutional database the mammographic reports for negative or benign cases with follow-up and the histological records when lesions were assessed by percutaneous biopsies and/or surgery. For each of the three subsets of 16 studies, the senior radiologist was asked to choose five easily assessable cases, five cases of intermediate difficulty, and 6 cases of higher interpretative difficulty to reflect the spectrum and complexity of the diagnoses encountered by the breast radiologists in the daily practice.

For all target studies, data both on breast density classification and on mammographic assessment according to BI-RADS (1) were collected, as reported in Table 2.

Cases with microcalcifications of interest with intermediate and high levels of perception and interpretative difficulty included studies with localized and clustered fine microcalcifications (BI-RADS 4), mostly low density and up to 1 cm in extent, and a few BI-RADS 2 cases, chosen among stable findings on previous studies or previously assessed as benign at percutaneous stereotactic biopsy. On the contrary, clustered microcalcifications of high density were chosen as easy cases. For true spiculated lesions versus negative cases, the perception and interpretative difficulties were due to the fact that many instances in both subsets were chosen among studies with ACR c and d density pattern and operated patients. To make the task more challenging, regional and diffuse microcalcifications (BI-RADS 2), therefore not of interest, were present in six negative cases.

All DBT studies were acquired using a Selenia Dimension 3D System mammography device (Hologic, Marlborough, MA), with a pixel pitch in the range 90–110 μm, depending on the selected reconstruction function. Exposure parameters were 27–36 kVp and 32–170 mAs with a 70 cm source to detector distance. The collected DBT studies were coded and blindly evaluated—according to a randomization scheme that implies, for each reader, the randomization of only cases within each of the investigated monitor—by one mid-experienced breast radiologist (junior-1, more than 2 years of experience in breast imaging) and one low-experienced breast radiologist (junior-2, a resident with 1 year of experience in our Unit), as well as by the senior radiologist. In addition, two independent interpretative sessions (first session, second session) were performed by the junior breast radiologists. A washout period of 2 weeks elapsed between each interpretative session. For the sake of consistency, the same software, SecurView from Hologic, was used on all the four types of displays for viewing the test images in all the interpretative sessions.

The study was conducted under standardized conditions in the same reading room and under a constantly subdued condition of ambient room light, less than 20 lux, to minimize reflections. To simulate the environment of routine clinical interpretation, any reader had, for each case, two paired monitors with the mammographic study and the associated DBT images (in some cases all the four standard projections, in others only the images acquired as workup of the native mammographic study) and was free to adjust the window level and width of image visualization as well as to use zoom function; no time constraint was imposed but overall time was recorded by the observers for each case. The observers were asked to complete an evaluation form for each of the 48 DBT studies, rating presence or absence of the aforementioned breast

<table>
<thead>
<tr>
<th>Kind of Breast Lesion</th>
<th>Microcalcification</th>
<th>Spiculated lesion</th>
<th>No Suspicious Lesion</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of lesions</td>
<td>18</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Age (years)</td>
<td>57 (48–76)</td>
<td>55 (36–69)</td>
<td>63 (48–73)</td>
</tr>
<tr>
<td>Breast thickness (mm)</td>
<td>49 (31–70)</td>
<td>58 (48–79)</td>
<td>50 (26–77)</td>
</tr>
<tr>
<td>BI-RADS assessment – 1</td>
<td>–</td>
<td>–</td>
<td>10</td>
</tr>
<tr>
<td>BI-RADS assessment – 2</td>
<td>7</td>
<td>–</td>
<td>6</td>
</tr>
<tr>
<td>BI-RADS assessment – 3</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>BI-RADS assessment – 4</td>
<td>9</td>
<td>10</td>
<td>–</td>
</tr>
<tr>
<td>BI-RADS assessment – 5</td>
<td>2</td>
<td>6</td>
<td>–</td>
</tr>
<tr>
<td>ACR breast density – a</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ACR breast density – b</td>
<td>4</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>ACR breast density – c</td>
<td>12</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>ACR breast density – d</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

ACR, American College of Radiology; BI-RADS, Breast Imaging Reporting And Data System.
lesions on a five-point confidence level (%CL): ≤ 20, 20–40, 40–60, 60–80, and >80. In addition, the localization of each finding was reported by breast laterality and quadrant, together with additional data regarding the overall quality of the images. The observers were unaware of the number of potential findings, so they were free to report more than one finding per case. Prior to any first session, a training session was held on the filling of the evaluation form.

Statistical Analysis

The results of the 1152 observations (48 images × 4 monitors × 3 observers × 2 sessions) were recorded in a database for statistical analysis.

First of all, a concordance study was performed to estimate the intra-/interobserver reproducibility as well as the reproducibility between monitors by jointly considering the type of lesion and its localization (breast laterality and quadrant) as a single pivotal variable. Due to the nature of this variable, the reproducibility was evaluated by computing the Cohen’s kappa statistic $k_c$ (13) that was interpreted in a qualitative manner on the basis of the Landis and Koch classification criteria (14). The observer-truth agreement was finally assessed by considering as truth the reference evaluation.

Subsequently, data were also analyzed by applying the well-defined statistical tools developed within the multireader-multicase (MRMC) scheme (15,16). The MRMC scheme is a fully crossed design in which all the observers (R) evaluate every image (I) with all monitors (M) under investigation, leading to a total of $R \times I \times M$ evaluations. To this end, a five-point rating scale was generated a posteriori, by opportune combining the data regarding the type of lesion and the CL of the observers’ evaluation, so that a score of 1 indicates high confidence that the patient is nondiseased (no lesion and CL > 80%) and a score of 5 indicates high confidence that the patient is diseased (lesion and CL > 80%). The same score was attributed to each localization, for cases presenting more than one localization. The developed five-point score and the reference evaluations were used to investigate the performance of each medical display and of each observer in terms of alternative free-response receiver operating characteristic analysis (AFROC). Specifically, the weighted jackknife AFROC (wJAFROC) (17) figure of merit (FoM) was computed by using the Obuchowski-Rockette procedure with Hillis improvement as significance-testing method (18). A weight of 1 or 0.5 was assigned to images showing one or two lesions, respectively (19). The trapezoidal area under the empirical AFROC curve was used as a nonparametric FoM to quantify observer performance by using the RJafroc R-package (available at https://cran.r-project.org/web/packages/RJafroc/index.html). The JAFROC paradigm was chosen to take into consideration both the detection and localization of the suspected abnormalities. Finally, the strength of association between the considered medical displays and the overall quality of the image was assessed by the Fisher exact test (20).

All statistical analyses were performed using SAS (version 9.2, SAS Institute Inc., Cary, NC) and R (version 3.3.0, R Foundation for Statistical Computing, Vienna, Austria) software.

RESULTS

Concordance Study

To assess the different patterns of reproducibility, we considered as concordant the cases in which the observers correctly identified the type of lesion and its localization. Cases showing two localizations were considered as concordant ones only if both the localizations were identified.

The intraobserver reproducibility (agreement between the two sessions) ranged from 0.920 to 1.000 and 0.751 to 0.897 for junior-1 and junior-2, respectively. An almost-perfect level of agreement was observed for all the considered medical displays when junior-1 was considered, with a $k_c$ value of 1 for monitor C; a similar pattern was observed for junior-2 for monitors A and D, whereas lower $k_c$ values for monitors B and C were obtained.

To assess the interobserver reproducibility and the between-monitor reproducibility, the data of the second session were used. As reported in Table 3, we observed an almost-perfect agreement when we compared junior-1 versus senior (range: 0.868–0.950), whereas lower values were obtained in the junior-2 versus senior comparison (range: 0.738–0.851); the agreement between junior-1 and -2 ranges from 0.794 to 0.850. Of note, the highest interobserver reproducibility values were obtained on monitor C and monitor D.

As reported in Table 4, by considering the observer-truth agreement an almost-perfect agreement was observed for monitors C and D for all the observers. For monitor A, we observed the lowest $k_c$ values (from 0.744 of junior-2 to 0.823 of senior).

| TABLE 3. Interobserver Reproducibility at Second Reading Session |
|-------------------|-------------------|-------------------|
|                   | Junior-1 vs Junior-2 | Junior-1 vs Senior | Junior-2 vs Senior |
| Monitor A         | 0.813 (0.668; 0.940) | 0.868 (0.759; 0.977) | 0.738 (0.593; 0.884) |
| Monitor B         | 0.794 (0.665; 0.923) | 0.871 (0.763; 0.978) | 0.797 (0.670; 0.924) |
| Monitor C         | 0.850 (0.741; 0.960) | 0.950 (0.883; 1.000) | 0.851 (0.742; 0.960) |
| Monitor D         | 0.849 (0.738; 0.961) | 0.876 (0.774; 0.978) | 0.825 (0.707; 0.943) |

In each cell is reported the $k_c$ value together with the 95% confidence interval (95% CI).
**Performance Evaluation**

Table 5 shows the individual wJAFROC values observed within each monitor for the three observers and the relative average value. A not statistically significant difference was observed between monitors C and D (P value: 0.170) and between monitors A and B (P value: 0.287). Performances with monitors A and B were significantly lower than those obtained with monitor C (A vs C: P value < 0.001; B vs C: P value < 0.001) and monitor D (A vs D: P value < 0.001; B vs D: P value: 0.002). These results are also confirmed by looking at the percentage (readers’ average) of correctly localized abnormalities in respect to that identified in the reference evaluation that ranges from 79.2% for monitor A to 95.8% for monitor C. As concerns the percentage of nonabnormalities correctly identified, the highest values were observed for monitors C and D (95.8%) followed by monitors A (91.7%) and B (89.6%).

**Additional Information**

Regarding the global quality of the DBT target images, according to our results the majority of them were scored by all the observers as optimal or excellent. It is interesting to note that, for the senior radiologist, most images were considered of optimal quality with monitors A and B, whereas almost all images were deemed excellent with monitors C and D. Very few images (maximum of two per observer) were considered of suboptimal quality. Junior-1 and senior radiologist scored more images as excellent with displays C and D with respect to displays A and B. Specifically, for each observer significant associations were obtained between the self-reported quality of the image and the displays (P values from Fisher’s exact test: <0.001 for junior-1, 0.015 for junior-2, <0.001 for senior radiologist).

Finally, as concerns the reading time, for displays A and B it ranged from 1 to 5 minutes, whereas for displays C and D the minimum reading time moves to 2 and 3 minutes, respectively. Also, for senior reporting, the median reading time moves from 2 minutes for displays A and B to 5 minutes for displays C and D, suggesting a possible difficulty in adapting to a 1000 cd/m² display luminance level, significantly different from that used for years.

**DISCUSSION**

With mammography, especially in screening programs, it is challenging to detect lesions of small size and/or low contrast with respect to the surrounding normal tissue. To be perceived, a lesion has to exceed the human physiological threshold both in size and in contrast and, as the human eye’s contrast sensitivity is an inverse function of the size of the target (21), small lesions need higher contrast to be seen. If a single microcalcification is detected by a single element on the digital capture device of 0.070 mm spatial resolution and its image is displayed on a 5 MP monitor of 0.165 mm, such as those conventionally used in mammography, the observer may not be able to detect it at the current display luminance level. So, either the observer can further magnify the lesion so it is large enough to be detected, or he/she can increase the display luminance level so that the microcalcification is more easily detectable by the human eye.

A previous study has shown that the perceived image quality increases as the luminance of the illuminator increases (21). In detail, the perceived image quality continues to improve up to 3000 cd/m² and then drops beyond 6000 cd/m² because of excessive glare.
In the present study, to evaluate the influence of display luminance level on the detectability of breast microcalcifications and spiculated lesions, 48 DBT studies have been selected in which the presence or the absence of the lesions were of easy, intermediate, or difficult perception, according to a high-experienced breast radiologist and regardless of the BI-RADS classification. All studies were displayed on four monitors of the same brand and independently evaluated by a senior radiologist and two junior breast radiologists with different levels of mammographic expertise. According to our results, the performances achieved with lower luminance monitors were significantly lower than those obtained with higher luminance monitors, thus confirming what was suggested by Kimpe and Xthona (9): increasing the calibrated luminance of a medical display increases the detection probability of breast lesions with DBT.

As regards false negative (FN) outputs (with respect to the lesion of interest), we observed cases (5, 4, and 1) on displays A, B, and D, respectively, whereas no FN outputs were observed on monitor C, by looking at the senior evaluation. A similar trend of FN evaluations was observed for junior-1 (7 and 0 FN with monitor A and monitor C, respectively) and junior-2 (9 and 4 FN with monitor A and monitor C, respectively). Interestingly, the majority of the aforementioned FN occurred on microcalcification cases, indicating that microcalcifications need higher luminance and contrast to be seen. Regarding spiculated lesions, our results did not show a dependence of detectability from display luminance level. These observations are not surprising, as DBT is known to easily depict radiated spicules, while fine microcalcifications with low radiopacity can sometimes be overlooked at DBT as well as with 2D mammography, more often with standard than synthetic mammography.

Finally, some notes about the possible limitations of the current study should be mentioned. Although our case series is too small to support any definite conclusion, we grasped as much as possible information from the data by resorting to two suitable statistical approaches to provide evidence about the role of luminance levels on the observer’s performance. Two additional issues regard the senior radiologist (who selected the cases and also participated in the observer’s performance study) and the mono-institutional nature of the study. We are aware that the inclusion of an observer (the senior one) that both selects and reads the cases could introduce a bias. However, according to both the overall number of cases evaluated by the senior radiologist during its routine clinical activity and the time interval that occurred between case selection and case evaluation (more than 2 months), we...
are confident in the generated results. Moreover, the data generated from this study, as well as those generated by our other substudies focused on different radiological modalities, could be viewed as the first step toward a deeper evaluation of the impact of the physical parameters of medical displays in the reporting process. By taking advantage of the information acquired within this first study, we are planning to set up additional studies involving larger cohorts of patients and observers as well as involving other Institutes to improve the transferability and generalizability of the observed results.

In conclusion, our study supports the importance of the display luminance level for the perception of challenging microcalcifications. Even though our study has a small number of investigated images and the tests were performed on the medical display devices of only one of the manufacturers producing high-luminance value panels, its results can be useful when a new mammographic system is under evaluation for purchase. Discussions about the diagnostic accuracy more often focus on the technical characteristics of mammographic detector equipment and the radiologists’ expertise. However, the displays for image visualization are not a negligible component: according to our study, a higher display luminance level can improve the percentage of correct detection for radiologists with different expertise.

REFERENCES