# Delta$^4$ Phantom+

## Pre Treatment QA for all machines

## TrueBeam, Versa HD, Radixact, Ethos
## TomoTherapy, Halcyon, Unity, MRIdian


**Testimonial**

*"The Delta$^4$ Phantom+ is a valuable tool for us at Herlev Hospital. We find great value in an independent measurement of the treatment plans. At Herlev Hospital we now have a complete range of Scandidos Delta$^4$ products covering TrueBeams, Halcyons and MRIdian."*

**Ulf Bjelkengren, Technical Manager, M.Sc.,**
**Medical Physics, Herlev and Gentofte Hospital**

QA from prescription to final fraction

**Delta$^4$ Family** one software platform for all your QA needs.

With Quality Assurance from prescription to final fraction you can now increase your workflow efficiency and be confident that the treatment dose delivered to your patient is safe.

Delta$^4$
by ScandiDos

*Innovative and Efficient QA*
*www.delta4family.com*

# A four-alternative forced choice (4AFC) methodology for evaluating microcalcification detection in clinical full-field digital mammography (FFDM) and digital breast tomosynthesis (DBT) systems using an inkjet-printed anthropomorphic phantom

Lynda C. Ikejimba[a)], Jesse Salad, Christian G. Graff, Bahaa Ghammraoui, and Wei-Chung Cheng
*US Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA*

Joseph Y. Lo
*Medical Physics Graduate Program, Duke University, 2424 Erwin Road, Durham, NC 27705, USA*

Stephen J. Glick
*US Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA*

(Received 11 September 2018; revised 12 April 2019; accepted for publication 26 April 2019; published 5 July 2019)

**Purpose:** The advent of three-dimensional breast imaging systems such as digital breast tomosynthesis (DBT) has great promise for improving the detection and diagnosis of breast cancer. With these new technologies comes an essential need for testing methods to assess the resultant image quality. Although randomized clinical trials are the gold standard for assessing image quality, phantom-based studies can provide a simpler and less burdensome approach. In this work, a complete framework is presented for task-based evaluation of microcalcification (MCs) detection performance for DBT imaging systems.
**Methods:** The framework consists of three parts. The first part is a realistic anthropomorphic physical breast phantom created through inkjet printing, with parchment paper and iodine-doped ink. The second is a method for inserting realistic MCs fabricated from calcium hydroxyapatite. The reproducibility and stability of the phantom materials were investigated through multiple samples of parchment and ink over 6 months. The final part is an analysis using a four-alternative forced choice (4AFC) reader study. To demonstrate the framework, a task-based 4AFC study was conducted using a clinical system to compare performance from DBT, synthetic mammography (SM), and full-field digital mammography (FFDM). Nine human observers read images containing MC clusters imaged with all three modalities and tried to correctly locate the MCs. The proportion correct (PC) was measured as the number of correctly detected clusters out of all trials.
**Results:** Overall, readers scored the highest with FFDM, (PC = 0.95 ± 0.03) then DBT (0.85 ± 0.04), and finally SM (0.44 ± 0.06). For the parchment and ink samples, the linear attenuation properties were very stable over 6 months. In addition, little difference was found between the various parchment and ink samples, indicating good reproducibility.
**Conclusions:** This framework presents a promising methodology for evaluating diagnostic task performance of clinical breast DBT systems. © *2019 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.13629]

Key words: anthropomorphic phantom, calcium hydroxyapatite, digital breast tomosynthesis, parchment paper

## 1. INTRODUCTION

Although screening mammography has contributed to the decrease in breast cancer mortality in the past 30 years,[1] it still has room for improvement, especially for women with dense breast tissue. In clinical trials, digital breast tomosynthesis (DBT) has been shown to improve the cancer detection rate and reduce the proportion of patients that are recalled for additional imaging studies, compared to full-field digital mammography (FFDM).[2,3] As of 2018, there are five DBT systems that have been approved by the U.S. Food and Drug Administration (FDA), and these DBT systems are very different in design. For example, they vary in acquisition geometries, detector hardware, and reconstruction methods. Additionally, vendors continue to modify their DBT systems, modifications which require regulatory approval to ensure safety and effectiveness. Although the gold standard for assessing performance of DBT systems is the randomized control clinical trial,[4,5] this process is expensive, time-consuming, and often involves administering additional radiation dose to the patient. Researchers in academia, industry, and government are developing methods for objectively assessing diagnostic task performance of breast imaging systems without conducting clinical trials.

One approach is the so-called "virtual clinical trial" (VCT),[6–8] where the entire imaging chain is modeled including a digital phantom, a model of the imaging system, and a model of the human observer reading the simulated imagery. Using a digital phantom, a group at the FDA has recently conducted a VCT named VICTRE (Virtual Imaging Clinical Trial for Regulatory Evaluation) that has emulated a clinical trial comparing FFDM and DBT that was used for regulatory approval.[9] Although VCTs have great potential to enable more efficient assessment of DBT image quality, they do involve complex modeling of the imaging system of interest.

To circumvent the difficulty of accurate modeling in a VCT, another approach is to use physical phantoms to emulate the breast. Physical phantoms are particularly useful since they can be used directly on the imaging system, thereby directly capturing the true hardware and software characteristics. Important phantoms for regular system quality control (QC) include the American College of Radiology (ACR) mammography phantom,[10] which has been FDA approved for clinical FFDM systems, and the phantom, used in Europe for similar purposes.[11] These phantoms contain signals in a uniform background, and they are very useful for routine QC. However, for system optimization or task-based evaluation, phantoms should possess anatomical texture and realism. The CIRS BR three-dimensional (3D) phantom (CIRS Inc., Norfolk, VA) is a physical phantom that is not based on patient anatomy, but instead mirrors the power law spectrum properties of the breast. In addition, some virtual phantoms were successfully realized and used in scientific studies, such as the UPENN[12] and Duke[13] physical phantoms. A number of researchers have demonstrated how task-performance-based image quality can be assessed using physical phantoms. Cockmartin et al.[14] describe a shakable phantom with acrylic spheres, used to evaluate FFDM and DBT performance with regard to detecting masses and microcalcifications (MCs). This novel phantom is part of a useful methodology to evaluate commercially available systems using a four-alternative forced choice (4AFC) scheme, but it is somewhat limited by material realism, as well as MC specks being in uniform immediate background. Another notable phantom is the so-called Doublet phantom,[13] used with inkjet-printed signals in a 4AFC scheme to compare FFDM, DBT, and synthetic mammography (SM) performance. This phantom, fabricated from a patient breast CT scan, was unique but had had a minor drawback in terms of resolution and material realism. Our group has recently developed a novel approach that uses inkjet printing to rapidly fabricate a physical phantom based on a virtual model.[15] When signals are included, this phantom can be utilized as part of a method for task-based evaluation of mammography systems.

This work has two main objectives. For the first objective, a novel methodology is presented for evaluating signal detection performance in mammographic imaging systems. The methodology consists of the following components. One component is a physical phantom comprising realistic materials, manufactured through rapid ink-jet printing. The next component is an insert with pathological features that is physically placed into the phantom and contained within the breast, consisting of MCs of various sizes. The final component is a reader study using a 4AFC scheme for measuring detection performance. For the second objective, the methodology is used on one vendor's system to compare detection performance for small MCs in three breast imaging modalities: DBT, conventional FFDM, and SM. Preliminary efforts were done to test the feasibility of this approach,[16] but the present work presents a substantial expansion of the study with a more robust approach and statistical analysis. This work thus presents a self-contained methodology to objectively assess task-based system performance of MC detection in FFDM, DBT, and SM.

## 2. METHODS

### 2.A. Digital breast phantom

The first step in fabricating the physical phantom is generating a digital breast phantom for printing. In theory, any digital breast phantom could be used; in this study, the method of Graff[17] was used to generate a procedural, rule-based digital anthropomorphic breast phantom. First, an analytic expression for an uncompressed breast surface was used to define the boundaries of the phantom. A 1-mm skin layer and a nipple were added to the anterior side of the surface and a backing muscle layer was added to the posterior side. In the interior of the surface, glandular compartments were defined based on a random Voronoi segmentation. For each segmented glandular compartment, a ductal tree was grown using a random branching algorithm starting from the nipple. At the terminal branches of each ductal tree, terminal duct lobular units were added. Initially, the interior of the phantom was assigned to be purely glandular tissue. Random fatty lobules were inserted to create subcutaneous and perilobular fat layers along with some fatty structures within the glandular regions. Each fatty lobule was incased by a ligament structure. Fat lobules were added until the desired glandular fraction of 28% was achieved. A glandular fraction of 28% was chosen to yield a sufficient amount of dense, challenging background regions for the study, while still maintaining a realistic overall breast density. Because only one phantom was printed for this study, it was important to have many areas with glandular structure. Finally, starting from the backing muscle layer, simulated arteries and veins were grown into the breast interior using an approach similar to the one used for duct generation. The breast tissue locations were determined on a 3D grid with equal voxel spacing. During the printing process, the breast model was binarized: all nonfatty tissues were binned to the glandular value. Because the model is virtual, it is possible to sample the volume at any resolution. An isotropic voxel size of 70 μm was selected based on the detector element spacing as well as the thickness of the paper onto which the model would be printed.

To simulate the compression of breast tissues during a mammographic examination, a finite element elasticity

model was employed using the FeBio software package (www.febio.org). A tetrahedral mesh version of the phantom was generated and fat or glandular elasticity properties were assigned to each mesh element based on the local tissue components. Two moving planar surfaces were used to compress the breast model during a dynamic linear elasticity simulation. The compression simulation ended once the target compressed breast thickness of 4 cm was achieved, representing approximately an average breast thickness. Slice numbers and fiducial markers were then added to the volume to assist with alignment of the physical phantom.

## 2.B. Physical breast phantom

After generating the digital breast phantom, the physical phantom was fabricated through inkjet printing. The process has been described in detail in previous work from this group,[15] but a summary is presented here. As mentioned, all tissue classes in the virtual model were binarized, so the physical phantom consisted of two materials: one to simulate fibroglandular tissue and another to simulate fat tissue. The virtual model was realized by printing with fibroglandular-mimicking ink onto fat-mimicking paper. A standard desktop Epson inkjet printer (Epson Workforce WF-3620, Long Beach, CA) was used. To simulate the fibroglandular tissue, customized ink was created by doping regular dye ink (InkThrift, Vermont PhotoInkjet, East Topsham Village, VT) with iohexol (Omnipaque, GE Healthcare, Princeton, NJ). The iohexol had an iodine concentration of 350 mg/ml, and the iodine-ink solution contained iohexol and ink at a ratio of 33% and 67%, respectively. The fat tissue was simulated using parchment paper (King Arthur, Norwich, VT), each sheet measuring 70 μm in thickness. The entire volume of the virtual phantom was fabricated in a slice-by-slice fashion as each slice of the model was printed onto a corresponding sheet of paper. The phantom was then constructed.

To assemble the sheets, holes were made over the fiducial markers using a custom hole punch to ensure alignment. Figure 1 provides a side-by-side comparison of a slice through the virtual model and the same slice printed on a sheet of parchment paper. The sheets were then placed onto a support plate which contained posts along the chest wall, corresponding to the positions of the holes in the sheets. Each sheet was slid over the posts and arranged in sequential order. About 600 sheets were carefully stacked in this manner. A cover plate containing holes for the posts was then placed on top to secure everything. While the section with the sheets was about 4 cm in thickness, the addition of the support and cover plates resulted in a total phantom thickness of about 6.5 cm.

To investigate the robustness of the phantom, test swatches were created with iodine-doped ink printed onto parchment paper and assessed in terms of their reproducibility and their stability over time. To create a test swatch, 30 squares were printed onto parchment paper using the iodine-doped ink. Each printed square measured 15 mm × 15 mm. The squares were then stacked one on top of the other to make a sample about 2.2 mm thick. An example is shown in Fig. 2.

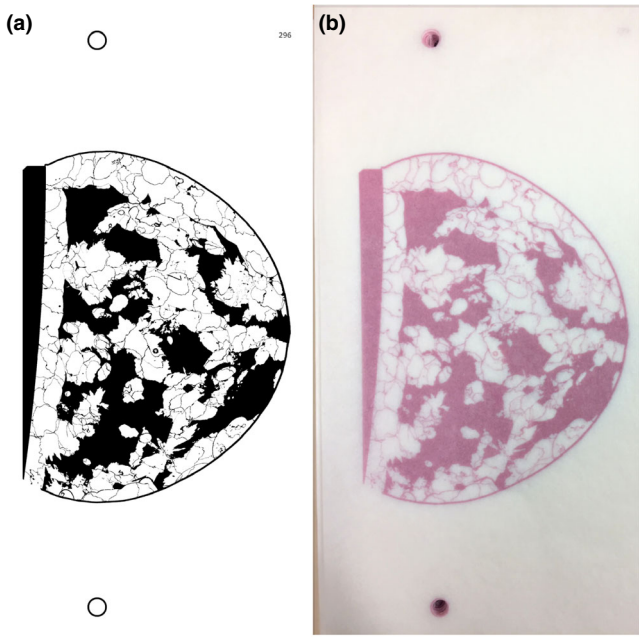


FIG. 1. Virtual and realized phantom. A side-by-side comparison is given of (a) a slice through the virtual model with fiducial markers visible (black circles) and (b) the corresponding printed parchment sheet with punched holes. The slice number can be seen. [Color figure can be viewed at wileyonlinelibrary.com]

Four different batches of ink were manufactured, each made several days apart; a given batch produced between three and five test swatches. The test swatches were then imaged on a polychromatic Hologic Lorad Selenia system with a Mo/Mo source and filter at 28 kVp and 10 mAs. Using the raw ("For Processing") images, the combined effective linear attenuation coefficient ($\mu$) of the parchment and iodine-doped ink was measured for each test swatch, and the average $\mu$ of each batch was investigated over 6 months. As a reference, the effective linear attenuation coefficient was measured from a CIRS tissue-equivalent chip approximating 100% fibroglandular tissue, measuring 20 mm × 20 mm × 5 mm. Statistical analysis was performed to determine if there was any significant difference between the batches or changes to attenuation over time.

## 2.C. MC template and inserts

Microcalcifications were included as pathological features and fabricated with real calcium hydroxyapatite (HA), primarily because it is associated with malignant lesions found in clinical breast cancer cases.[18] To simulate the MCs, disks of HA were made from commercially available raw HA powder (Lot # MKBX3842V, Sigma-Aldrich, St. Louis, MO, USA) via mechanical pressing.[19] A given amount of HA was mixed with 1% (weigh ratio) polyvinylpyrrolidine (PVP) K30 in a tumbling mixer (W.A. Bachofen, Basel, Switzerland) for 10 min at 50 rpm. The mixture was then granulated to a 1:0.2 weight ratio of powder to binder mixture, with 3% (weigh to volume ratio) PVP alcoholic solution. The obtained

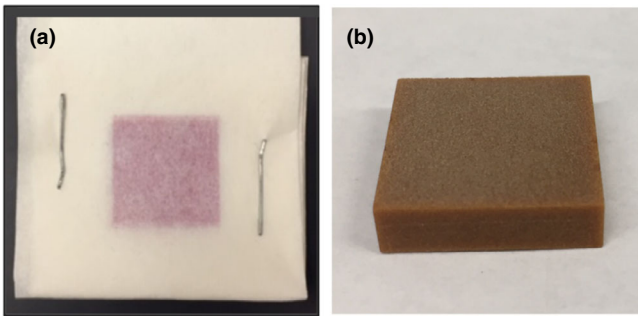


FIG. 2. Iodine and parchment test swatch. (a) Test swatches were created by printing a 15 mm × 15 mm square using iodine-doped ink onto parchment paper. Thirty squares were stacked to make a 2.2-mm thick sample. (b) A tissue-equivalent chip simulated 100% fibroglandular tissue. [Color figure can be viewed at wileyonlinelibrary.com]

substance was vacuum dried at 30°C for 24 h, followed by compression into flat-faced 8-mm diameter disks using a rotary tableting machine (PICCOLA, Riva S.A., Buenos Aires, Argentina). The weight and thickness of the disks were adjusted to yield approximately the same density of 1.9 g/cm$^3$, using 20 kN target compression load for each disk.

To create the inserts, the tablets were first crushed using a mortar and pestle, shown in Fig. 3, then sieved to separate specks into a size range of 150–180 μm. A stencil was created to contain the specks by poking holes in a pentagon arrangement into a sheet of Mylar. The stencil contained a five-by-five array of these pentagons, spanning 10 mm across the widest points and spaced 20 mm apart from center to center. The Mylar stencil was affixed onto double-sided tape and the HA specks were individually placed into the holes and secured with another sheet of the double-sided tape on top. The tape-stencil-tape combination was then sealed with sheets of parchment paper on either side. The insert measured approximately 500 μm in total thickness. To assist with extracting regions of interest (ROIs), 2 mm #102 nipple markers from Y-SPOT® (Beekley Medical, Bristol, CT, USA) were placed as fiducial markers onto the insert, at known positions outside the area containing the MC clusters. For imaging, the completed insert was placed within the center of the phantom stack, as shown in Fig. 4. Because each sheet of the phantom had a printed slice number, the insert could be placed at the same height above the detector for all acquisitions.

## 2.D. Image acquisitions

Images were acquired on a clinical Hologic Selenia Dimensions (Bedford, MA, USA) system under HD combo mode, allowing for acquisition of FFDM, DBT, and SM data in rapid succession. The source to imager distance was 700 mm. A direct detection a-Se detector was used with an imager size of 4096 rows by 3328 columns and a pixel pitch of 70 μm. The settings under automatic exposure control (AEC) were used. The average glandular dose (AGD) was similar across all three modalities. Summarized imaging parameters are presented in Table I.

The AEC settings for FFDM specified a target/filter combination of W/Rh, with 50 μm Rh thickness, and a tube setting of 32 kVp and 206 mAs, resulting in an AGD of approximately 2.6 mGy. Both processed and unprocessed FFDM projections were obtained, but only the processed ("For Presentation") images were used in the study. No pixel binning was performed. For DBT, the AEC used W/Al (700 μm Al) with 34 kVp and 56 mAs, yielding an AGD of about 2.4 mGy. The detector type were the same as FFDM, but 2 × 2 pixel binning was performed, resulting in a pixel pitch of 140 μm. Fifteen projections were taken over a 15° angular span at approximately 1° increments. The DBT volumes were reconstructed using filtered back projection. SM images were generated from the DBT data through the vendor's proprietary algorithm. The resulting in-plane reconstructed resolution was about 100 μm for both SM images and reconstructed DBT volumes.

To obtain signal present images, the insert was placed in the middle of the phantom and scanned three times. Between each scan, the insert was slightly shifted in a random x–y direction to obtain a new position and yield more background samples. The insert remained in the same z-position for all acquisitions. For signal absent ROIs, a single scan was taken of the phantom without the insert; additional scans were not needed, as they would yield essentially the same image.

## 2.E. ROI extraction

Code was written in MATLAB (Version R2017a, Mathworks, Natick, MA, USA) to extract the ROIs containing the MC clusters. Using the fiducial markers, the position and angle of the insert can be automatically determined from the radiographic images, and the ROIs can be extracted with the

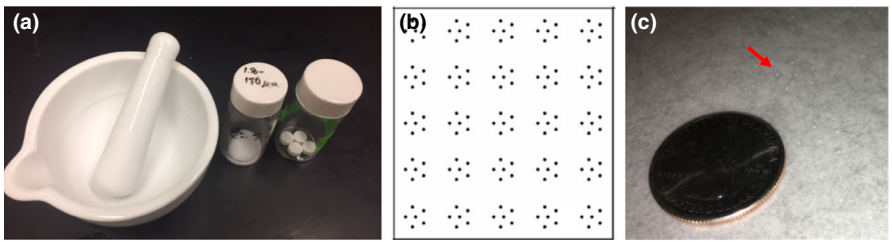


FIG. 3. Fabrication and placement. (a) The HA was pressed into tablets and crushed with a mortar and pestle. (b) The pentagon-shaped clusters were arranged in a five-by-five grid over several background locations. (c) The MC clusters are shown with a quarter for reference. The red arrow annotates a pentagon cluster. [Color figure can be viewed at wileyonlinelibrary.com]

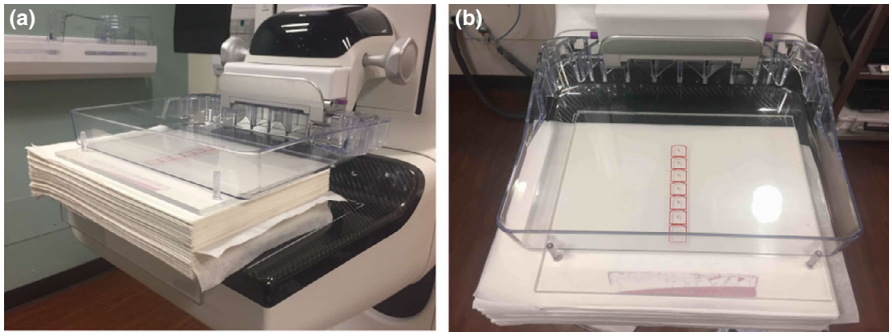


FIG. 4. MC insert with phantom. The phantom with MC insert is shown on the clinical system from the (a) side and (b) top. [Color figure can be viewed at wile yonlinelibrary.com]

TABLE I. Summary of imaging acquisition parameters.

| Modality | Target/Filter | kVp | mAs | AGD (mGy) |
|---|---|---|---|---|
| FFDM | W/Rh | 32 | 200 | 2.6 |
| DBT | W/Al | 34 | 54 | 2.4 |
| SM | W/Al | 34 | 54 | 2.4 |

AGD, average glandular dose; AEC, automatic exposure control; DBT, digital breast tomosynthesis; FFDM, full-field digital mammography; SM, synthetic mammography.
Beam conditions were determined by AEC settings.

TABLE II. Specifications of ROIs.

| Modality | Signal present ROIs | In-plane pixel size (μm) | ROI sizein mm | ROI sizein pixels |
|---|---|---|---|---|
| FFDM | 47 | 65 | 20 × 20 | 304 × 304 |
| DBT | 49 | 105 | 20 × 20 × 9 | 183 × 183 × 9 |
| SM | 47 | 105 | 20 × 20 | 183 × 183 |

DBT, digital breast tomosynthesis; FFDM, full-field digital mammography; ROIs, regions of interest; SM, synthetic mammography.
ROIs were extracted using custom software.

cluster relatively centered. Figure 5 shows the phantom with the BBs automatically detected in [Fig. 5(a)], and with markers placed over their centroids in Fig. 5(b). A summary of the extraction parameters is given in Table II.

From the three signal present acquisitions, a total of 47 or 49 unique signal present ROIs were extracted per modality. Of these, ten ROIs were used for reader training, and the remaining 37 or 39 ROIs were used for testing. The extracted ROIs are shown in Fig. 6 for an FFDM acquisition. The extracted ROIs measured 20 mm × 20 mm, resulting in a window size of 183 pixels × 183 pixels or 304 pixels × 304 pixels depending on the resolution of the modality. Digital breast tomosynthesis volumes of interest contained nine slices, with the center slice corresponding roughly to the slice in which MCs were most in focus. From the signal absent acquisition, about 450 ROIs were extracted with partial overlap, from which 111–117 ROIs (three times the total number of signals) were randomly selected for the study. This ensured that each signal absent ROI was viewed only once in the study.

## 2.F. 4AFC reader study

A task-based reader study was performed using a 4AFC scheme, facilitated with the publicly available program Foursquares.[20] For the study, nine nonradiologist readers participated, each familiar with the study and types of images. Prior to the study, the readers underwent supervised training to become familiar with the study protocol and software interface. Readers were instructed to sit at a comfortable distance from the monitor and to maintain this viewing distance

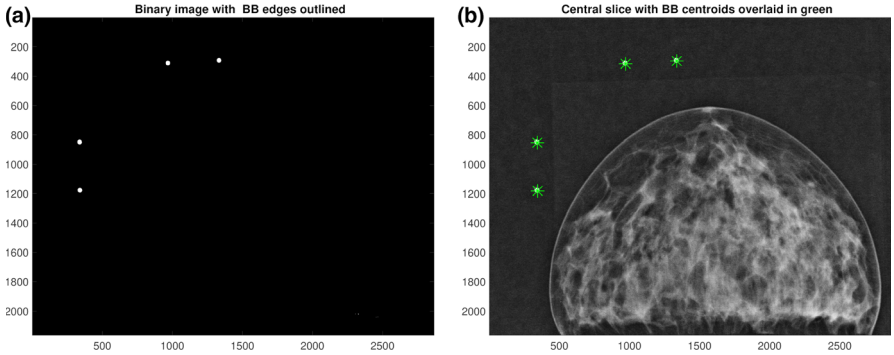


FIG. 5. Fiducial Markers. Fiducial markers are automatically detected in the image (a), and centroids are indicated with green asterisks (b). [Color figure can be viewed at wileyonlinelibrary.com]

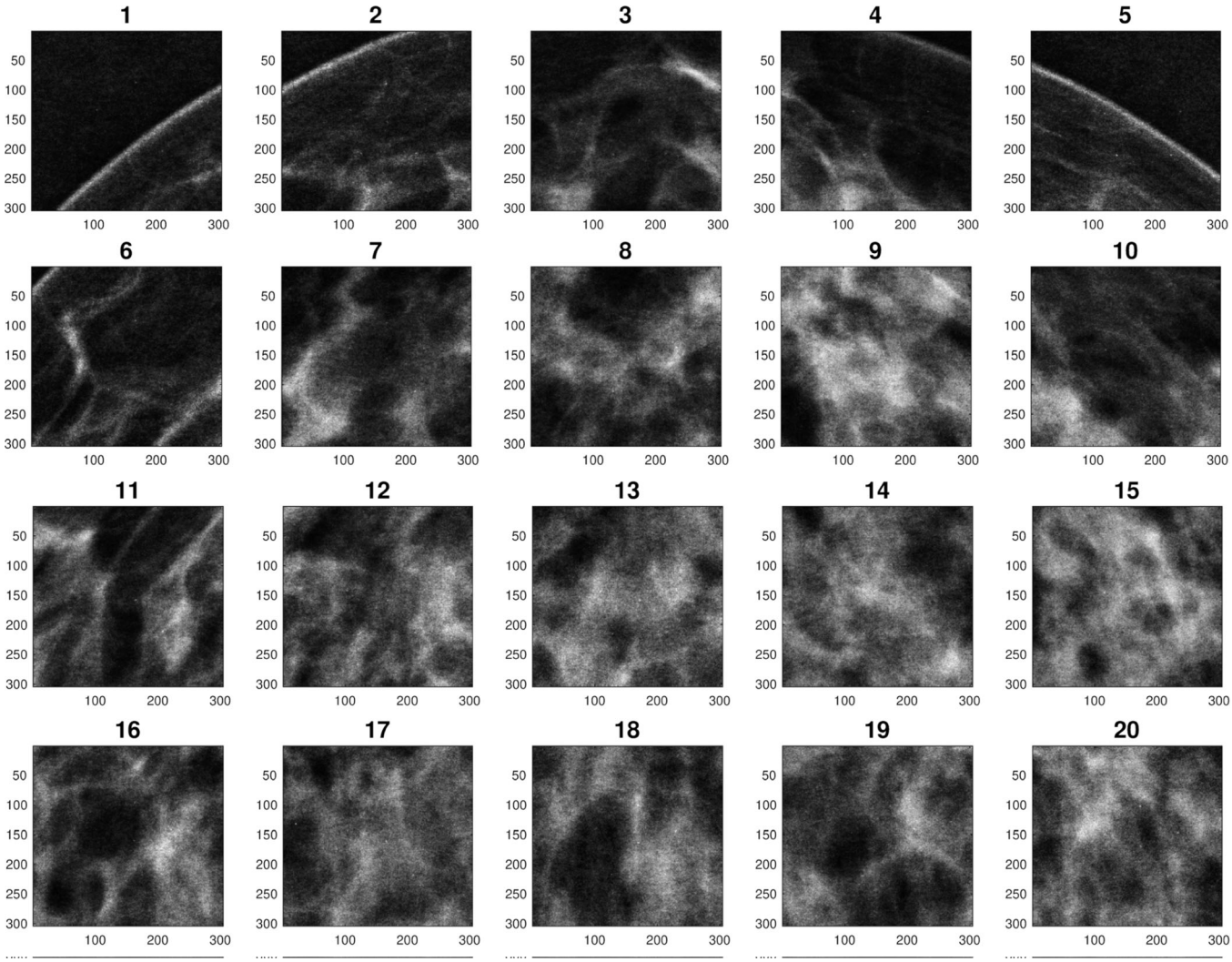


FIG. 6. Extracted ROIs. Extracted ROIs containing MC clusters are shown for an FFDM acquisition. ROIs containing the breast boundary were excluded from the study. FFDM, full-field digital mammography; ROIs, regions of interests.

throughout the study. Readers were not allowed to change the display contrast or magnification. During training, four images were loaded in the program, only one of which contained a true positive. To the right of the program window, a cue signal was presented with a MC cluster in a uniform background to illustrate the task. An example is presented in Fig. 7. Readers were instructed that the cue signal would demonstrate the approximate size and contrast level of the test signal, but the exact location and orientation of the task could vary. Training was done with ten ROIs from each modality, and immediate feedback was provided to the reader indicating whether the selection was correct or incorrect.

The images were displayed with a 30", 6-Mega pixel medical display (Coronis Fusion 6MP DL MDCC-6130, Barco NV, Kortrijk, Belgium). The 26" by 16" active screen area carries 3,280 × 2,048 color pixels. The display was set to operate in the Diagnostic mode with the grayscale standard display function (GSDF), 300 cd/m$^2$ maximum luminance, and 6500K white point. The DICOM GSDF was verified with a spectroradiometer (CS-1000A, Konica Minolta Sensing Americas, Inc., Ramsey, NJ, USA) and passed the AAPM TG18 10% acceptance test. The visual experiments were conducted in a dark room designed for human vision studies. The lighting was controlled to emulate a radiology reading room. Ambient light was emitted from the ceiling and did not create glares on the display. The desktop background of the monitor was in dark gray, and the illuminance measured from the display was 2.71 lux. ROIs were displayed at a 1:1 magnification. For each ROI, a default window/level was found by first trying to read the window center and width from the DICOM header. If that was not available or the data were invalid, a window was set by the maximum and minimum pixel values, and the level was set to be the mean pixel value.

Statistical analysis of the results was performed using iMRMC,[21] a freely available software package developed within this division (https://github.com/DIDSR/iMRMC/releases). Scores were computed as the proportion correct (PC), the number of images a reader correctly located out of all cases. The PC provides a measure of reader performance directly the 4AFC data.[14,22,23]

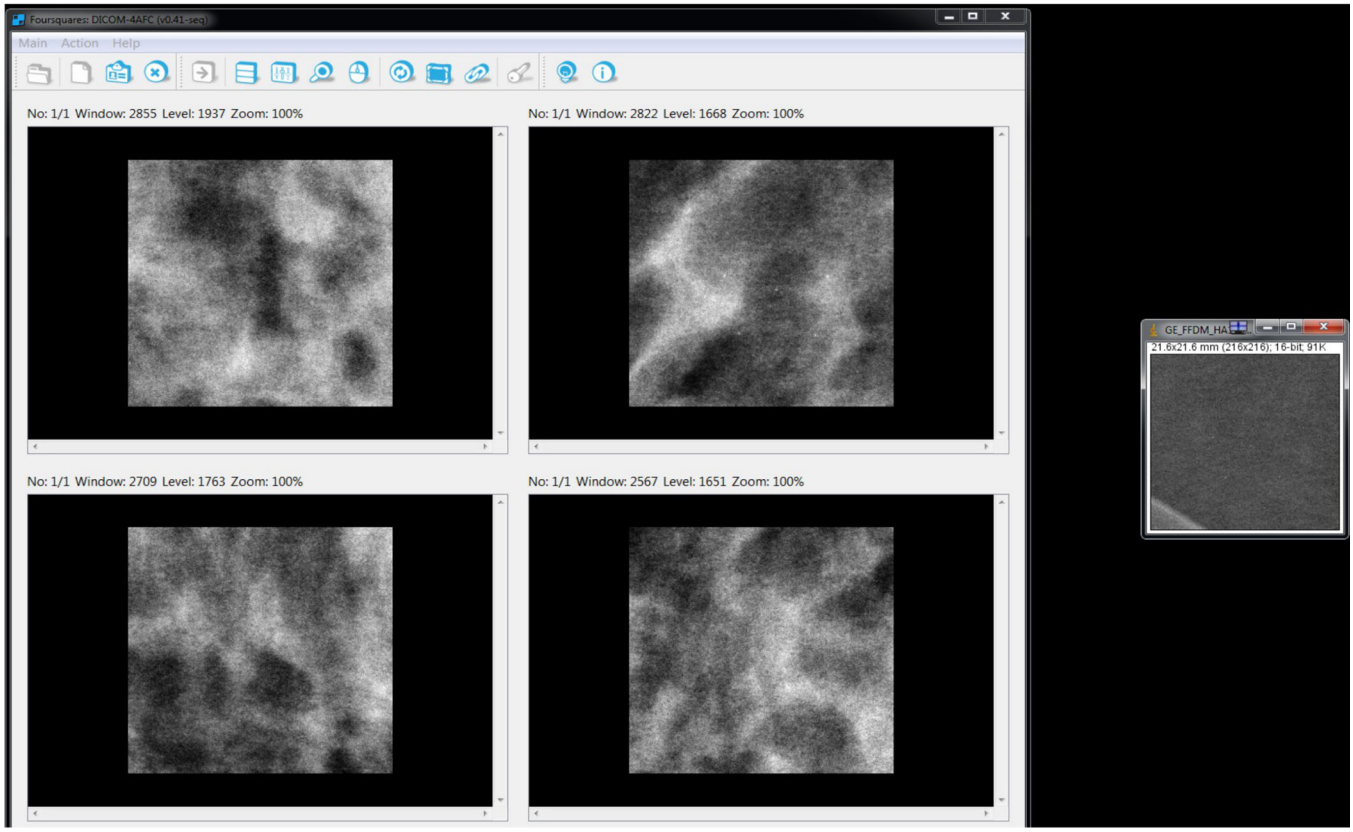


FIG. 7. 4AFC study. the display that is presented to the reader from the Foursquares 4AFC program is shown. A cue ROI (right window) was presented to indicate the approximate intensity of the MCs. 4AFC, four-alternative forced choice; MCs, microcalcifications; ROI, regions of interest. [Color figure can be viewed at wileyonlinelibrary.com]

## 3. RESULTS

Measurements are presented in Fig. 8 for values of the effective linear attenuation coefficient $\mu$ taken over 6 months. Here, the $\mu$ values are plotted over time for the four batches of iodine-doped ink and parchment test swatches as well as for the CIRS fibroglandular-equivalent chip. Each data point for the batches represents the average of all test swatches. Error bars were smaller than the line data markers.

Overall, the $\mu$ values of the iodine and parchment test swatches were between 0.11 $mm^{-1}$ and 0.12 $mm^{-1}$, compared to the reference value of 0.11 $mm^{-1}$ for the CIRS fibroglandular-equivalent chip. A two-way ANOVA was performed to determine if either time or the batch number had a significant effect on the $\mu$ values, treating time as a fixed effect and batch number as a random effect. No statistical significance was found in the change in attenuation across months ($P > 0.05$). The difference between batches was found to be significant ($P < 0.01$). However, the actual inter-batch variability was very small; the largest difference was about 0.01 $mm^{-1}$, and the values across different batches were generally within 10% of each other.

The reader-averaged PC scores for all three modalities are given in Fig. 9 as a box plot. The red center line indicates the median score, and the bottom and top blue edges represent the 25$^{th}$ and 75$^{th}$ percentile scores, respectively. The whiskers

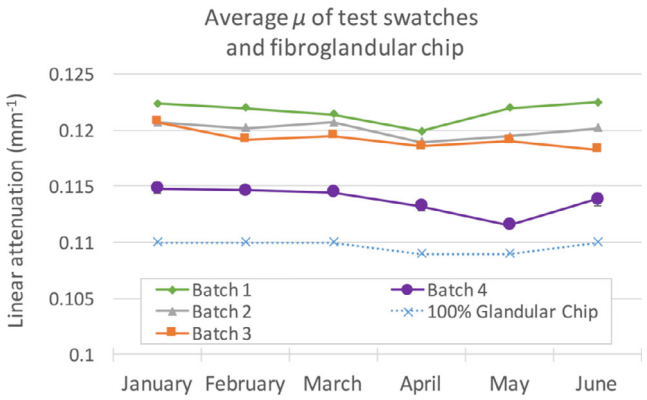


FIG. 8. Phantom stability over time. The effective linear attenuation coefficients from the ink and parchment test samples are plotted over 6 months, along with the reference values for fibroglandular tissue. Values were stable over time [Color figure can be viewed at wileyonlinelibrary.com]

mark the most extreme data not considered an outlier. For DBT and FFDM, scores were clustered around the maxima, so no upper whisker is show. In addition, the FFDM results were almost bimodal with the exception of one reader. As a result, the 75$^{th}$ percentile and maxima were the same, the bottom whisker was the single unique value, and the median was the average of the two modes.

The average performance was observed to vary with modality. The scores were highest for FFDM, where the PC

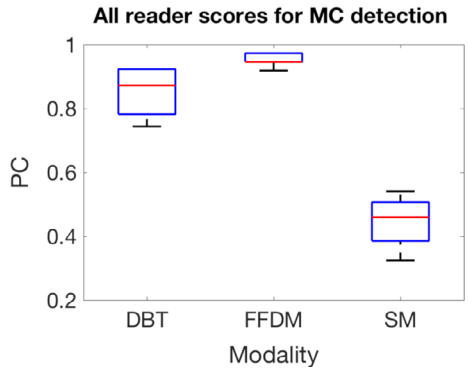


FIG. 9. Reader-averaged PC. The PC was averaged across all readers for each imaging system. Performance with SM was significantly lower than with DBT or FFDM. DBT, digital breast tomosynthesis; FFDM, full-field digital mammography; PC, proportion correct; SM, synthetic mammography. [Color figure can be viewed at wileyonlinelibrary.com]

was 0.95 ± 0.03. The next highest scores were seen for DBT, where average PC was 0.85 ± 0.04. Finally, scores were lowest for SM, where the PC was 0.44 ± 0.06 for SM. The difference in PC between FFDM and DBT was found to be statistically significant, with p = 0.02. The performance with SM was significantly lower than that of both DBT ($P << 0.01$) and FFDM ($P << 0.01$).

## 4. DISCUSSION

Using an anthropomorphic breast phantom is useful for system comparisons, but inserting an MC model into the phantom can be challenging. One approach that has been used with the CDMAM phantom[11] is to emulate signals with very thin gold disks. Although this approach provides small objects with mammography contrast emulating that of similar size MCs, the objects are gold and have different attenuation coefficient than clinical MCs. Furthermore, the use of thin gold disks is questionable for use in assessing DBT image quality because they are anisotropic, being much larger in the x–y direction than the z-direction. Cockmartin et al.[14] have described a clever "shakable" anthropomorphic phantom with background consisting of PMMA spheres of various shapes placed into water. In that phantom, MC signals were modeled by gluing specks of calcium carbonate (CaCO3) to a PMMA plate of thickness 2 mm. The plate is then inserted into the phantom at a specified depth. While novel, one disadvantage of this approach is that a typical 1 mm DBT slice containing the calcifications could have a large contribution from the uniform PMMA background. Thus, in that particular slice, the phantom would appear to be less anthropomorphic. For the phantom presented here, the calcium hydroxyapatite MCs are placed within a template that is on the order of 500 μm thick, and thus would contain more of the nonuniform background within a 1-mm DBT slice. This template provides a useful way to introduce signals into a breast phantom for task-based assessment.

The 4AFC reader scores suggested that the detection of small MCs is highest with FFDM, followed by DBT, with a sharp drop-off in performance with SM. Such results may be explained by the differences in the acquisition process. For example, the Selenia Dimensions FFDM system uses an amorphous-selenium detector with detector pixels of size 70 μm. The DBT acquisition protocol uses 2 × 2 binning of the projection data before reconstruction, effectively making the pixel size equal to 140 μm and thereby reducing spatial resolution. This system also uses a continuous x-ray tube motion during the scan, introducing a spatial blurring into the projection data. Finally, an antiscatter grid is in place during FFDM imaging, but this grid is removed for DBT imaging. Imaging without an antiscatter grid could potentially decrease the accuracy in detecting MCs. It should be pointed out that this study focused solely on the performance with the Hologic Selenia Dimensions system; performance for each of the three modalities could be very different for other breast imaging systems with different acquisition geometries, detectors, reconstruction methods, etc.

The rather substantial degradation in performance for detecting small MCs with SM as compared to FFDM and DBT has been observed in other studies. For example, Nelson et al.[24] compared image quality resulting with FFDM and C-view using both the ACR phantom and a novel 3D-printed anthropomorphic breast phantom. They concluded that C-view improved conspicuity of certain higher contrast objects, but overall provided poorer spatial resolution and noise properties, making it more difficult to visualize smaller MCs. MacKenzie et al.[25] compared the detection of simulated masses and MCs of various sizes, and concluded that the threshold diameter (i.e., the signal diameter needed to meet a prespecified level of detection accuracy) of MCs was considerably higher with SM. Although these studies suggest a decrease in performance for detecting small MCs with SM, the clinical significance of this finding is unclear. Assessing the benefits of screening mammography is a controversial topic, and one opinion is that the detection of small MCs contributes to overdiagnosis and overtreatment of breast cancer. This discussion is beyond the scope of this paper.

The task-based evaluation methodology described here provides a realistic modeling of a common diagnostic task in breast imaging: the detection of an MC cluster containing small individual MCs. In addition, the reproducibility of the phantom was tested by assessing material characteristics systems over time. Analysis of the $\mu$ values suggested that there were some differences between the various batches. However, all differences were relatively small, less than 10% of the mean signal, indicating that method is fairly reproducible. Although this approach appears to accurately mimic detection of MCs in clinical breast imaging, there are a number of approximations that could be improved on.

First, the phantom is designed to model x-ray attenuation of fat using parchment paper and fibroglandular tissue using iodine-doped ink printed onto parchment paper. A previous study using spectral analysis with a high-purity Germanium photon counting detector showed that the linear attenuation coefficients of these materials were close to their reference values (see Fig. 5 from Ikejimba et al.[15]). This study also showed that parchment paper from different companies had

different characteristics. It was found that parchment from King Arthur Bakery (Norwich, VT) was a good compromise between size of sheets sold and attenuation properties. However, the ink was doped with iohexol (concentration of iodine of 350 mg/ml). One problem with iodine is the strong K-edge at 33 keV. Thus, it would not be recommended to fabricate phantoms for use in analyzing mammography systems with a peak voltage than 33 kVp. In addition, the physical breast phantom used in this study was printed based on a binary digital breast phantom, so the only two tissue types modeled were fibroglandular and fat tissues. It would be possible to print other tissue types with varying attenuation characteristics by printing with varying gray levels. This is the approach that has been previously used in fabricating paper phantoms for CT.[26,27]

Another issue is with the parchment paper. Because the parchment paper is not perfectly uniform, a stack of parchment paper will yield an image with a slightly mottled appearance. Although this nonuniformity is not controllable, we believe that this feature adds a very low-level, realistic looking nonuniform random structure to the mammogram and DBT projections. Also, the square shape of the phantom is not anthropomorphic, since plain parchment paper is located outside of the breast support. This probably causes an unrealistic increase of scatter from paper outside of the breast. It might be feasible to cut the excess paper from each slice outside of the breast boundary, and we are currently investigating solutions for this.

Finally, there were some limitations with the MC placement. All the specks were placed in a single plane, while clinical MC clusters have an irregular 3D arrangement. A new template is currently being developed that creates a more random arrangement of specks. In addition, there was a lack of MC search in the 4AFC study. While including a search task may not change the rank ordering of the modalities, the increased difficulty would likely lower overall performance for all conditions.

This work presents a practical approach to objectively assessing task-based performance of breast imaging systems. The methodology is useful in many ways. For regulatory purposes, the approach described herein can assess the safety and effectiveness of a system using an objective, quantitative evaluation. Alternative methods, like the ACR mammography phantom, only provide subjective assessments of image quality. Additionally, this methodology has relevant clinical implications for system optimization, such as determining the optimal acquisition parameters.[28,29] Previous system optimization efforts have used an assumption of uniform background, which can alter results.[30] Finally, if computer model observers are developed, for example, using a deep learning model observer,[31] this approach could be used for QC testing of breast imaging systems.

## 5. CONCLUSIONS

This work demonstrates the use of a novel, realistic breast phantom for assessing task-based performance in mammographic systems. Small, challenging MCs were fabricated and inserted into the phantom and imaged on a Hologic Selenia Dimensions system using FFDM, DBT, and SM modalities. A 4AFC study was conducted with nine human observers, and reader performance for MC detection was evaluated for each modality. Overall, readers scored the highest with FFDM, (PC = 0.95 $\pm$ 0.03) then DBT (0.85 $\pm$ 0.04), and finally SM (0.44 $\pm$ 0.06), with statistically significant differences between all scores.

With all the above components, this work provides a methodology for assessing image quality based on realistic diagnostic tasks. The methodology has a number of possible applications such as (a) improving assessment of safety and effectiveness in regulatory applications, (b) helping optimize system and design parameters for maximizing performance, and (c) quality control testing to assure maximum clinical performance over time. Future work will include modeling of extended masses into the phantom, and evaluation of performance using other breast imaging systems.

## ACKNOWLEDGMENTS

The authors acknowledge the help of Dr. Guo Zhang with the Foursquare software, and the help of Dr. Frank Samuelson and Dr. Brandon Gallas with the statistical analysis. This work was supported by a Critical Path grant from the Center for Devices and Radiological Health, with a fellowship administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

## CONFLICTS OF INTEREST

The authors have no relevant conflicts of interest to disclose.

[a)]Author to whom correspondence should be addressed. Electronic mail: Lynda.Ikejimba@fda.hhs.gov.

## REFERENCES

1. DeSantis CE, Fedewa SA, Goding SA, Kramer JL, Smith RA, Jemal A. Breast cancer statistics, 2015: convergence of incidence rates between black and white women. *CA: Cancer J Clin*. 2016;66:31–42.
2. McDonald ES, Oustimov A, Weinstein SP, Synnestvedt MB, Schnall M, Conant EF. Effectiveness of digital breast tomosynthesis compared with digital mammography: outcomes analysis from 3 years of breast cancer screening. *JAMA oncol*. 2016;2:737–743.
3. 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.
4. Kopans DB. Digital Breast Tomosynthesis From Concept to Clinical Care. *Am J Roentgenol*. 2014;202:299–308.

5. Smith R, Duffy S, Gabe R, Tabár L, Yen AMF, Chen THH. The randomized trials of breast cancer screening: what have we learned? *Radiol Clin North Am*. 2004;42:793–806.
6. Bakic PR, Barufaldi B, Higginbotham D, et al. Virtual clinical trial of lesion detection in digital mammography and digital breast tomosynthesis. *Proc. SPIE*. 2018;10573:13.
7. Barufaldi B, Higginbotham D, Bakic PR, Maidment AD. OpenVCT: a GPU-accelerated virtual clinical trial pipeline for mammography and digital breast tomosynthesis. *Proc. SPIE*. 2018;10573:1057358.
8. Elangovan P, Warren LM, Mackenzie A, et al. Development and validation of a modelling framework for simulating 2D-mammography and breast tomosynthesis images. *Phys Med Biol*. 2014;59:4275.
9. Badano A, Badal A, Glick S, Graff CG, Samuelson F, Sharma D, Zeng R. In silico imaging clinical trials for regulatory evaluation: initial considerations for VICTRE, a demonstration study, in *Medical Imaging 2017: Physics of Medical Imaging, Vol. 10132* (International Society for Optics and Photonics, 2017), pp. 1013220
10. American College of Radiolog.Mammography Quality Control Manual, Medical physicist's section, 225-330 (1999)
11. Bijkerk K, Lindeijer J, Thijssen M. The CDMAM-Phantom: a contrast-detail phantom specifically for mammography. *Radiology*. 1993;185:395.
12. Carton AK, Bakic P, Ullberg C, Derand H, Maidment AD. Development of a physical 3D anthropomorphic breast phantom. *Med Phys*. 2011;38:891–896.
13. Ikejimba LC, Glick SJ, Choudhury KR, Samei E, Lo JY. Assessing task performance in FFDM, DBT, and synthetic mammography using uniform and anthropomorphic physical phantoms. *Med Phys*. 2016;43:5593–5602.
14. Cockmartin L, Marshall NW, Zhang G, et al. Design and application of a structured phantom for detection performance comparison between breast tomosynthesis and digital mammography. *Phys Med Biol*. 2017;62:758.
15. Ikejimba LC, Graff CG, Rosenthal S, et al. A novel physical anthropomorphic breast phantom for 2D and 3D x-ray imaging. *Med Phys*. 2017;44:407–416.
16. Ikejimba LC, Yan T, Kemp K, et al. Methodology for the objective assessment of lesion detection performance with breast tomosynthesis and digital mammography using a physical anthropomorphic phantom. *Proc. SPIE*. 2018;10573:105735G.
17. Graff CG. A new open-source multi-modality digital breast phantom, Proc. SPIE, 978309-978309-978310 (2016)
18. Warren L, Mackenzie A, Dance D, Young K. Comparison of the x-ray attenuation properties of breast calcifications, aluminium, hydroxyapatite and calcium oxalate. *Phys Med Biol*. 2013;58:N103.
19. Ghammraoui B, Makeev A, Glick S. Classification of breast microcalcifications using dual-energy mammography. *Proc. SPIE*. 2018;10573:1057305.
20. Zhang G, Cockmartin L, Bosmans H. A four-alternative forced choice (4AFC) software for observer performance evaluation in radiology. *Proc. SPIE*. 2016;97871E.
21. Obuchowski NA, Gallas BD, Hillis SL. Multi-reader ROC studies with split-plot designs: a comparison of statistical methods. *Academic Radiology*. 2012;19:1508–1517.
22. Han M, Jang H, Baek J. Evaluation of human observer performance on lesion detectability in single-slice and multislice dedicated breast cone beam CT images with breast anatomical background. *Med Phys*. 2018;45:5385–5396.
23. Timberg P, Båth M, Andersson I, Mattsson S, Tingberg A, Ruschin M. Visibility of microcalcification clusters and masses in breast tomosynthesis image volumes and digital mammography: a 4AFC human observer study. *Med Phys*. 2012;39:2431–2437.
24. Nelson J, Wells J, Samei E. TU-CD-207-08: intrinsic image quality comparison of synthesized 2-D and FFDM images. *Med Phys*. 2015;42:3611–3612.
25. Mackenzie A, Kaur S, Elangovan P, Dance DR, Young KC. Comparison of synthetic 2D images with planar and tomosynthesis imaging of the breast using a virtual clinical trial. *Proc. SPIE*. 2018;10577:105770H.
26. Jahnke P, Limberg FR, Gerbl A, et al. Radiopaque three-dimensional printing: a method to create realistic CT phantoms. *Radiology*. 2016;282:569–575.
27. Salad J, Ikejimba LC, Makeev A, Graff CG, Ghammraoui B, Glick SJ. Development of a physical anthropomorphic breast phantom for objective task-based assessment of dedicated breast CT systems. *Proc. IWBI*. 2018;10718:107180S.
28. Chan H-P, Goodsitt MM, Helvie MA, et al. Digital breast tomosynthesis: observer performance of clustered microcalcification detection on breast phantom images acquired with an experimental system using variable scan angles, angular increments, and number of projection views. *Radiology*. 2014;273:675–685.
29. Zeng R, Park S, Bakic P, Myers KJ. Evaluating the sensitivity of the optimization of acquisition geometry to the choice of reconstruction algorithm in digital breast tomosynthesis through a simulation study. *Phys Med Biol*. 2015;60:1259–1288.
30. Berglund J, Johansson H, Lundqvist M, Cederström B, Fredenberg E. Energy weighting improves dose efficiency in clinical practice: implementation on a spectral photon-counting mammography system. *J Med Imaging*. 2014;1:031003.
31. Alnowami M, Mills G, Awis M, et al. A deep learning model observer for use in alterative forced choice virtual clinical trials. *Proc. SPIE*. 2018;10577:105770Q.