Assessment of task-based performance from five clinical DBT systems using an anthropomorphic breast phantom

Lynda C. Ikejimba[a] Jesse Salad and Christian G. Graff
US Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA

Mitchell Goodsitt and Heang-Ping Chan
Michigan Medicine, University of Michigan, 1500 East Medical Center Drive, Ann Arbor, MI 48109, USA

Hailiang Huang and Wei Zhao
Stony Brook Medicine, Stony Brook University, 101 Nicolls Road, Stony Brook, NY 11794, USA

Bahaa Ghammraoui
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 27 April 2020; revised 7 September 2020; accepted for publication 18 October 2020; published 28 January 2021)

Purpose: Digital breast tomosynthesis (DBT) is a limited-angle tomographic breast imaging modality that can be used for breast cancer screening in conjunction with full-field digital mammography (FFDM) or synthetic mammography (SM). Currently, there are five commercial DBT systems that have been approved by the U.S. FDA for breast cancer screening, all varying greatly in design and imaging protocol. Because the systems are different in technical specifications, there is a need for a quantitative approach for assessing them. In this study, the DBT systems are assessed using a novel methodology with an inkjet-printed anthropomorphic phantom and four alternative forced choice (4AFC) study scheme.
Method: A breast phantom was fabricated using inkjet printing and parchment paper. The phantom contained 5-mm spiculated masses fabricated with potassium iodide (KI)-doped ink and microcalcifications (MCs) made with calcium hydroxyapatite. Images of the phantom were acquired on all five systems with DBT, FFDM, and SM modalities where available using beam settings under automatic exposure control. A 4AFC study was conducted to assess reader performance with each signal under each modality. Statistical analysis was performed on the data to determine proportion correct (PC), standard deviations, and levels of significance.
Results: For masses, overall detection was highest with DBT. The difference in PC was statistically significant between DBT and SM for most systems. A relationship was observed between increasing PC and greater gantry span. For MCs, performance was highest with DBT and FFDM compared to SM. The difference between PC of DBT and PC of SM was statistically significant for all manufacturers.
Conclusions: This methodology represents a novel approach for evaluating systems. This study is the first of its kind to use an inkjet-printed anthropomorphic phantom with realistic signals to assess performance of clinical DBT imaging systems. *© 2020 American Association of Physicists in Medicine. This article has been contributed to by US Government employees and their work is in the public domain in the USA.* [https://doi.org/10.1002/mp.14568]

Key words: digital breast tomosynthesis, iodine, mass, microcalcification, synthetic mammography

1. INTRODUCTION

The advent of screening mammography has been one of the driving forces that has resulted in a 39% reduction in breast cancer mortality.[1] Nonetheless, sensitivity and specificity of mammography are limited in certain population groups, most notably women with dense breast tissue and women on hormone replacement therapy.[2] One of the challenges with conventional full field digital mammography (FFDM) is the structural tissue overlap that can inherently obscure diagnostic features in two-dimensional (2D) imaging of the breast. Digital breast tomosynthesis (DBT) is a limited angle tomographic breast imaging modality designed to reduce the superposition of breast tissue, commercially introduced in 2011 for combined usage with conventional 2D mammography. As of 2020, statistics from the Mammography Quality Standards Act showed 69% of certified facilities in the United States offered DBT.[3] Recently, synthetic mammography (SM), a method designed to generate a mammography-like image from the DBT stack of image slices, has been introduced with the goal of reducing radiation by eliminating the standard 2D mammography acquisition.[4]

As of 2019, five commercial DBT systems have been approved by the U.S. Food and Drug Administration (FDA). The GE Senographe Essential (SenoClaire) (GE Healthcare, Waukesha, WI) has been approved for screening with DBT + SM, the GE Senographe Pristina (GE Healthcare, Waukesha, WI) for DBT + SM, the Hologic Selenia Dimensions (Hologic, Bedford, MA) with DBT + FFDM or DBT + SM, the Fuji ASPIRE Cristalle (Fujifilm, Stamford, CT) system for DBT + FFDM or DBT + SM, and the Siemens MAMMOMAT Inspiration (Siemens, Erlangen, Germany) for FFDM + DBT and DBT alone. Although these systems all perform tomosynthesis, they have many design and operational differences including acquisition geometry, exposure techniques, x-ray tube target and filter, detector type, use of varying reconstruction method, and different levels of radiation dose to the breast. Although clinical trials have demonstrated that DBT can improve breast cancer detection while reducing the false-positive recall rate,[5–7] it is unclear how clinical performance depends on the particular DBT system used, since to date no appropriately powered clinical studies have been conducted to compare different commercial DBT systems. Unfortunately, this type of comparison study is difficult to perform due to the high cost and complexity of conducting such a clinical trial. To circumvent some of the limitations of clinical studies, phantom-based methodologies are being developed for system evaluation.

Previous studies have described the development of breast phantoms and methodologies to assess imaging systems, and in general, the approaches can be broadly classified as virtual or physical. In virtual clinical trials (VCTs), each component of the imaging chain is simulated, including the breast, the imaging system, and the reader.[8–10] Virtual clinical trials are becoming more common for assessing new technology. Recently, a research group at the FDA conducted the VICTRE (Virtual Imaging Clinical Trial for Regulatory Evaluation) trial demonstrating the use of VCTs in a regulatory application.[8] While such approaches can be efficient and allow for a large number of subjects, they require accuracy in modeling the imaging components. As a result, modeling clinical systems for use with VCTs can prove challenging when processing software is proprietary. The other approach to assessing system performance is with use of physical phantoms. For quality control (QC), standard phantoms include the American College of Radiology (ACR) phantom[11] and CDMAM phantom,[12] approved for accreditation purposes in the United States and Europe, respectively. While these phantoms are ideal for quick or routine QC testing, they contain signals in a uniform background and thus may not be sufficient for optimization studies, where system performance can change with anatomical complexity.[13] Physical structured phantoms exist for system optimization such as the Penn[14] and Duke[13] phantoms, both based on anatomical properties and fabricated through additive manufacturing. In addition, the phantom described by Cockmartin et al.[15] consists of acrylic spheres of varying sizes in a water bath. Masses and microcalcifications (MCs) can be added for task-based assessment of mammography systems. Although these phantoms are very useful and unique in their approach, the materials used can be somewhat limited in realism. There is a need for a realistic, anthropomorphic physical breast phantom that can be used for QC, system optimization, and regulatory evaluation of system effectiveness.

Our research group at the FDA has previously developed a methodology to objectively assess task performance of breast x-ray imaging systems using a realistic anthropomorphic breast phantom.[16] This phantom uses a novel inkjet printing approach to fabricate a physical phantom based on a virtual breast model. In addition, diagnostic features such as realistic MC clusters and extended masses can be inserted into the phantom. Region of interest (ROI) and volume of interest (VOI) images containing diagnostic features can then be extracted for use in performance assessment studies. A previous proceedings paper[17] from our group described very preliminary work. The present submission represents a greatly expanded study with substantive changes. The present paper contains five more figures, two additional tables, and a more rigorous statistical analysis. In addition, the present paper includes a substantial additional reader study evaluating detection of MC clusters. For this, we have used a novel method for fabricating MCs, as well as an approach for generating a template modeling random MC clusters that was inserted into the 3D paper phantom. This study evaluating MC detection was not included in the conference proceedings.

The goal of this study was twofold. First, we endeavored to explore whether this task-based phantom assessment methodology was feasible for use on each currently available commercial DBT system. Each commercial system uses proprietary image processing software, and it is known that imaging of some phantoms produces image artifacts that would not occur during imaging of patients.[18] The second goal was to use this task-based assessment methodology to compare performance achieved with the five FDA-approved commercial DBT systems under DBT, FFDM, and synthetic mammography (when available) imaging modes. Phantom images from each system were acquired using the automatic exposure control settings for that system. Thus, all phantom acquisition settings, and subsequent radiation dose levels, were dictated from the manufacturer settings. The results of this comparison could provide insight into the different tradeoffs associated with varying operational and design strategies used with different commercial DBT systems.

2. MATERIALS AND METHODS

2.A. Breast phantom fabrication

The breast phantom used in the study was a custom-made, 3D parchment paper phantom fabricated using an inkjet printing process described in detail previously.[16,19] A digital breast phantom was first created through analytical modeling by making a shell for skin, then dividing the interior into fibroglandular and adipose compartments, and finally, adding ligaments and blood vessels.[20] The digital phantom modeled a breast with 28% fibroglandular density, representing a dense breast with an extensive parenchymal structure. The digital phantom was then compressed to 4 cm thickness modeling the administration of the breast compression paddle. The model was sampled at 70-μm isotropic voxel resolution to match the thickness of the paper onto which it would be printed. Fiducial markers were inserted into every slice of the digital phantom to assist with proper registration of the printed sheets. The fiducial markers were designed as rings placed on the medial and lateral sides of the breast, separated by a fixed distance from each other and from the chest wall. These can be seen in Fig. 1.

Inkjet printing was used to realize the digital phantom. For the fibroglandular tissue, a custom ink solution was formulated by combining a ratio of 2/3 dye ink to 1/3 iohexol with an iodine concentration of 350 mg/mL. The final fibroglandular ink had an iodine concentration of 117 mg/mL. For the fat tissue, parchment paper served as the background onto which the ink for fibroglandular regions was printed. Printing was done on an Epson WF-3620 inkjet printer (Epson America, Long Beach, CA) with refillable ink cartridges. Before printing, each channel was assessed to ensure it could print a single color without "contamination" from other channels. To do this, a line pair pattern was printed with a single-color channel onto a parchment paper which, being slightly hydrophobic, would allow visualization of individual ink droplets. The samples were then examined under either a 5× optical microscope or jeweler's loupe. No droplets were observed from other color channels within each line or along the line edges, where color mixing would be most evident. This was repeated for every photographic setting available on the printer ("Photo glossy," "High quality," "Medium quality," etc.), totaling about 20 different settings. Printing proceeded only with a printer setting where no mixing was observed that also provided the greatest print speed. For a 4-cm compressed breast, a total of 571 sheets were printed. To house the sheets, a custom-made container was designed consisting of a 6-mm sheet of acrylic as a base; two posts extending vertically from the base and measuring roughly 6 mm in diameter and 76 mm in height, placed close to the chest wall; and an additional 6-mm thick sheet of acrylic on top to provide minor compression. The diameter of each post was set to match the diameter of the fiducial marker ring.

As previously mentioned, the fiducial markers were printed with each slice to ensure proper registration of the sheets. A hole was punched in each sheet through the center of the ring, using a custom hole punch designed for this purpose. The posts were then passed through the holes as each sheet was stacked. The holes were visually inspected and tested for fit on each sheet before proceeding to the next sheet.

2.B. Insertion of masses

Masses were fabricated in a similar inkjet printing manner. A three-dimensional mass was first digitally created using an implementation of lesion modeling software.[21] These masses had an approximate diameter of 5 mm with spiculations emanating outward from the center [see Fig. 1(a)]. Select parameters used in the mass generation included *number of initial segments* = 1358, *maximum number of neighborhood segments* = 0.98, and *mean radius decrease* of 0.89 as described in the work by de Sisternes et al.[21] The mass was duplicated to create an insert consisting of 18 identical masses arranged in a grid with 20 mm spacing center-to-center in the x- and y-directions, centered within the same z-slice. The mass insert spanned 70 pages in the

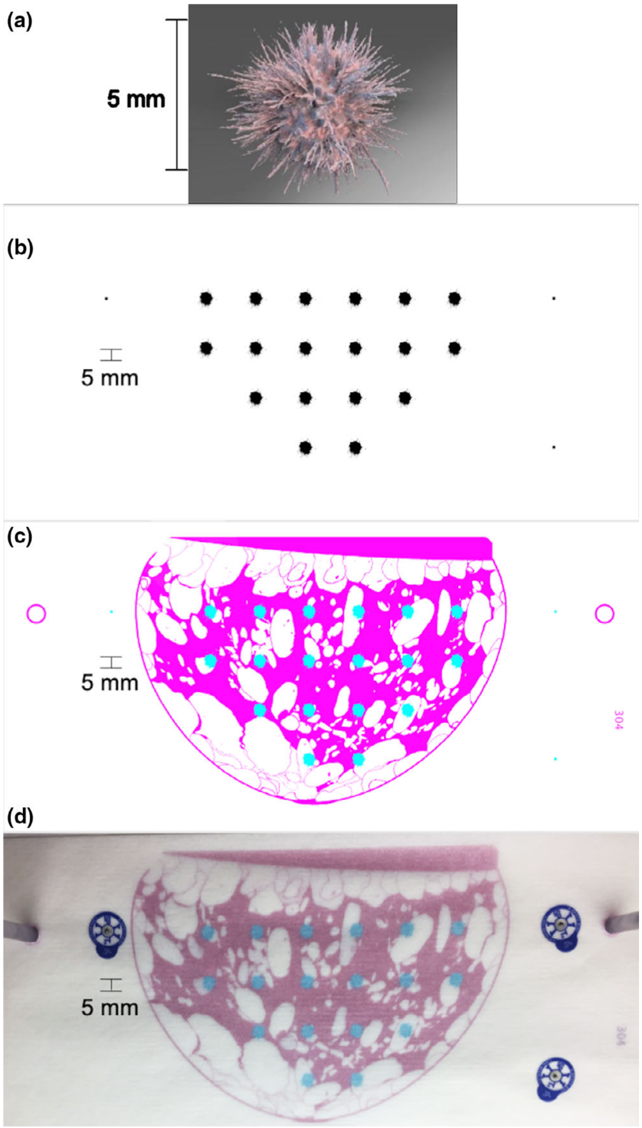

Fig 1. Insertion and printing of masses. (a) The three-dimensional lesion model is shown with spiculations. (b) The mass was duplicated and arranged into a grid with BB markers, with a two-dimensional central slice shown. (c) The grid was inserted into the virtual breast with ring fiducial markers and (d) printed with the corresponding type of ink. Portions of figure reprinted with permission from Ikejimba et al. "Assessment of task-based performance from five clinical Digital breast tomosynthesis systems using an anthropomorphic breast phantom," *15th International Workshop on Breast Imaging (IWBI2020)*. Vol. 11513 (2020). [Color figure can be viewed at wileyonlinelibrary.com]

z-direction, as the thickness of each sheet was 70 μm and the masses were 5 mm in diameter. This design maximized the number of ROIs that could fit within the breast while allowing sufficient space between each mass. The central slice of the mass insert contained three markers for BBs, used for automating ROI and VOI extraction. This insert was placed via pixel substitution into the central portion of the digital breast phantom, the region with the greatest area. To obtain more samples, four sets of mass inserts were created by shifting the entire grid of masses and BB markers in unison by 2–5 mm in the x- and y-directions, remaining centered in the same slice. As a result, a total of 72 ROIs containing a mass were created with unique background locations.

To print the masses, a new type of ink was synthesized. When higher concentrations of iohexol were used in the ink, the print heads were prone to clogging. To reduce clogging, a different ink/iodine solution was required to achieve a sufficiently high iodine concentration for the masses. A saltwater solution was made by dissolving potassium iodide (KI) in water at a concentration of 300 mg/mL. This was then mixed with ink at a ratio of 2/3 KI saltwater to 1/3 dye ink. The resulting ink for the masses had a KI concentration of 200 mg/mL. The mixture was placed in a separate color cartridge in the printer, different from the one containing the iohexol-based ink. Prior to printing, the fibroglandular and mass tissues were recolored in the digital model to match the cartridge color they would be printed with. This was done using GIMP (GIMP v 2.10.10, http://gimp.org), a freely available image manipulation program. The program allowed selected pixel values to be printed from a specific cartridge, enabling simultaneous printing of the two tissue types. Each slice containing the mass was printed onto a sheet of paper, yielding a subsection stack with 70 sheets of paper roughly 5 mm thick in total. The four 70-sheet mass inserts were each printed, producing four physical stacks of mass inserts. Figure 1 shows a side-by-side comparison of (a) the 3D lesion model, (b) the arrangement of mass lesions positioned within the breast with the three markers for BB locations, (c) the center slice of the breast phantom with inserted masses and the fiducial rings for sheet registration, and (d) a printed sheet of the same slice with the BBs in place and posts through the rings.

2.C. Microcalcifications

A new MC insert was created using an improved method similar to one previously described.[19] Using MATLAB, a template was first designed to mark locations where clusters would be placed. The template consisted of 5-row by 9-column grid of 5-mm diameter circles, each with an MC cluster. Within each cluster, the locations of the specks were randomly generated with a buffer around each speck to ensure they are placed within the 5-mm circle and do not overlap. The specks were made by combining calcium hydroxyapatite (HA) powder with the binding agent polyvinylpyrrolidone and compressing the mixture into a tablet using a mechanical press. The tablets were then crushed and separated by size using differential sieving. The resulting specks ranged in size between 150 and 180 μm. Five specks were placed into each of the pre-designated locations in the 5 mm circle, described above. The clusters were spaced 15 mm apart in x- and y-directions, forming a total of 45 clusters. Fiducial markers were included in the template to assist with extraction. The insert was sealed with a double-sided tape and more parchment paper. The excess paper and tape were trimmed to allow the insert to fit within the breast boundary of the printed phantom. The process of fabricating and inserting the MC template can be seen in Fig. 2.

2.D. Image acquisitions

Images of the phantom were acquired on five commercially available DBT systems: Hologic Selenia Dimensions at Sibley Memorial Hospital in Washington, DC; GE Senographe Essential (SenoClaire) at University of North Carolina-Chapel Hill (UNC) in Chapel Hill, NC; GE Senographe Pristina at University of Michigan in Ann Arbor, MI; Siemens MAMMOMAT Inspiration at the State University of New York (SUNY) in Stonybrook, NY; and Fujifilm Aspire Cristalle in Stamford, CT. Imaging was performed with FFDM, DBT, and SM modalities, with the exception of the GE Senographe Essential which lacked SM capability. The technical specifications of the systems are provided in Table I.

The imaging parameters were determined by the beam conditions used under automatic exposure control (AEC) for

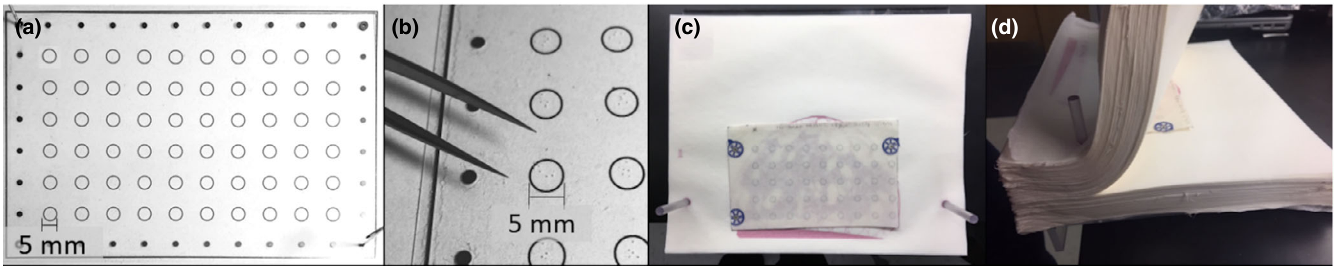

Fig 2. Fabrication and insertion of MC template. Clusters were made by manually placing MC specks within a 5-mm diameter circle, shown (a) from above and (b) as a close-up with visible specks. The completed template with BBs was placed between the central sheets of the printed phantom, shown (c) from above and (d) from the side. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE I. Summary of technical specifications.

	Hologic Selenia Dimensions	GE Senographe Essential (Sen oClaire)	GE Senographe Pristina	Siemens MAMMOMAT Inspiration	Fuji Aspire Cristalle
Version	AWS: 1.8.3.63, Cview: 2.0.1.1	Application ADS_56.21.3, RECON_01.10.4	Recon 02.8.7, SM: 2.3.0	VB60E\VX20A SL21P21 syngo VH22B SL19P26	FDR-3000AWS Mainsoft V9.0
Detector conversion	Direct	Indirect	Indirect	Direct	Direct
Anti-scatter grid in DBT	No	Yes	Yes	No	No
Detector version	CM862326	PLC0096_05	PXA0045_02	L03-00010	
Field of view (mm)[a]	217 x 267	239 x 306	239 x 285	238 x 299	236 x 296
Detector element size (μm)	70[b]	100	100	85	50[c]
In-plane pixel size (μm)	FFDM: 65 DBT, SM: 105[d]	FFDM, DBT: 100	FFDM, DBT:100, SM: 100	FFDM, DBT: 85, SM: 89	FFDM: 50, DBT, SM: 100
X-ray tube target	W	Mo or Rh	Mo or Rh	W	W
X-ray tube filtration	Al or Rh	Mo or Rh	Mo or Ag	Rh	Al or Rh
X-ray tube motion	Continuous	Step and shoot	Step and shoot	Continuous	Continuous
Angular range (deg)	15	25	25	50[e]	15
Number of projections	15	9	9	25	15
Source-to-imager distance (mm)	700	660	660	650	650
Reconstruction method	FBP	Iterative	Iterative	FBP	FBP

[a] Field of view in the FFDM acquisition.

[b] DBT uses 2x2 pixel binning.

[c] Pixels are hexagonal.

[d] In-plane resolution changes with slice number. This is the pixel size in the plane of focus.

[e] While gantry span is 50 degrees, acquisitions take place over 46 degrees.

each system. As a result, the x-ray tube settings ranged from 29 to 34 kVp for tube voltage, 40 to 180 mAs for total mAs for all projections, and 1.2 to 2.3 mGy for average glandular dose (AGD) as reported from the system display. A summary of the acquisition parameters is given in Table II. Note that these are the dose levels reported by the vendor. Because manufacturers may use very different assumptions for their calculation of displayed AGD, the system-displayed doses might have limited comparability. An independent dose calculation was performed using the IEC recommended[22] procedure with the Dance method.[23–25] Using the beam settings, the AGD was calculated to a reference phantom of 40 mm PMMA with the equation

$$D_T = K_E gcsT$$

where K_E is the entrance surface kerma; g, c, and s are factors dependent on the target, filter, and half value layer; and T is variable for commercially available DBT systems. This calculation is provided in the last column under Ref. AGD.

To determine the appropriate beam parameters, an image of the phantom was taken under AEC conditions. The phantom was positioned with its chest wall extended off the edge of the detector cover, and posts flush with the compression paddle, as illustrated in Fig. 3. This configuration was necessary to accommodate the height of the posts. Once the beam parameters were determined on each system, imaging was performed with the phantom repositioned with the overhanging lip against the detector cover and the compression paddle sitting atop the posts. The acquisition parameters were manually set and used to image the phantom with and without signals.

To image the masses, a central subsection of the phantom was replaced by the 70-sheet stack with printed masses. All four 70-sheet stacks containing masses were inserted and imaged one after the other in this manner on all systems, with the exception of the GE Essential since the fourth stack was not completed at the time of imaging. To image the MCs, the MC insert was placed between the central two slices of the phantom. Multiple acquisitions were taken in order to sample different background locations where the MC insert was shifted in a random manner between shots. For the signal-absent data, a single shot was taken of the phantom sheets without the inserted masses or calcifications. Additional scans were not necessary since the phantom background would remain the same.

ROIs and VOIs were extracted automatically from the images using a custom MATLAB program. Background ROIs were randomly selected from within the breast volume. This was achieved by extracting overlapping ROIs within the breast boundary in a raster fashion, producing on average between 350 and 450 ROIs. From these, enough background ROIs were randomly selected to equal 3× the number of signal present ROIs for a case cohort, considered all the cases for a given vendor, modality, and signal (e.g., Hologic FFDM masses). With this method, it is highly unlikely that an exact background ROI would be selected that corresponded to any signal present locations. To prevent learning the backgrounds, each signal absent ROI was randomly rotated by 90, 180, or 270 degrees. Mass ROIs were also randomly rotated by 90, 180, or 270 degrees and had an additional search component, whereby the center of the lesion could be located anywhere

TABLE II. Summary of acquisition parameters.

Vendor	Modality	Gantry Span	Target/filter	Tube Voltage (kVp)	Current-time (mAs)	x-y Voxel Size (μm)	AGD (mGy)	Ref. AGD (mGy)
Hologic Selenia Dimensions	DBT[a]/SM	15°	W/Al	32	65	105	2.1	1.66
	FFDM		WRh	30	180	65	2.0	1.63
GE Senographe Essential (SenoClaire)	DBT	25°	Rh/Rh	29	71	100	1.4	0.89
	FFDM		Rh/Rh	29	71	100	1.4	0.90
GE Senographe Pristina	DBT/SM	25°	Rh/Ag	34	40	100	1.5	1.08
	FFDM		Rh/Ag	34	40	100	1.5	1.09
Siemens MAMMOMAT Inspiration	DBT/SM	50°	W/Rh	30	200	85	2.3	1.86
	FFDM		W/Rh	30	100	85	1.2	0.93
Fuji Aspire Cristalle	DBT[a]/SM	15°	W/Al	33	52	100	1.9	1.47
	FFDM		W/Rh	30	89	50	1.2	0.84

[a] Detector uses 2x2 binning in DBT mode.

within a 10 mm × 10 mm area within the middle of the ROI. For the MCs, ROIs were rejected if the cluster was outside of the breast or too close to the breast boundary. The ROIs measured 15 mm × 15 mm for MCs and 20 mm × 20 mm for masses. DBT VOIs consisted of nine reconstructed DBT slices of 1 mm thickness. For MCs, between 92 and 106 ROIs were extracted per modality per system, and for masses between 44 and 62 ROIs.

2.E. Reader study

A 4AFC reader study was conducted to evaluate the detection of masses and MCs with each modality. Reading was conducted in a dark room designed for human reader studies, with the lighting in the room kept low to model that of a clinical reading room. Reading was performed on a 30" 6MP Coronis Fusion (6MP DL MDCC-6130, Barco NV, Kortrijk, Belgium) medical display calibrated to DICOM grayscale standard display function. The display contained an active screen area of 26" × 16" with 3280 × 2048 pixels. The display was operated in Diagnostic mode grayscale standard display function (GSDF) with 300 cd/m^2 maximum luminance. Ambient light from the ceiling did not cause glare on the monitor. The illuminance from the display was measured to be 2.71 lux. ROIs were displayed at a 1:1 magnification.

Scoring was performed by seven non-radiologist readers familiar with the type of images. Readers were medical physicists experienced in the given tasks with the 3D paper phantom. Reading was facilitated by the Foursquares software.[26] The program consists of four windows, each displaying an ROI or VOI. Only one of the windows contains a signal-present ROI or VOI, and the three others contained background images only. It was the objective of the reader to select the correct window. A "cue" image was presented next to the Foursquares program with a mass or one MC cluster in a uniform background; this provided the reader with an example of the true signal. For each experimental condition, ROIs within the case cohort were presented in a randomized order with respect to location within the breast phantom. As previously mentioned, a case cohort consisted of all the ROIs for a given vendor, modality, and signal — for example, 62 ROIs for Hologic DBT masses. The task for mass detection was a signal location unknown, and readers were instructed to perform some search. The ROIs for the masses were 20 mm × 20 mm, and the center of the mass could be located anywhere within the central 10 mm × 10 mm area of the ROI.

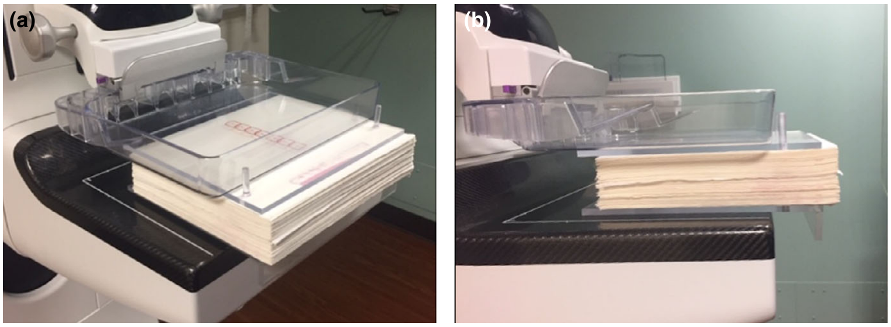

FIG 3. Phantom positioning during automatic exposure control (AEC) imaging. The phantom is positioned with the posts outside the field of view to allow for AEC estimation with accurate phantom height. The placement can be seen from the (a) top and (b) side views. [Color figure can be viewed at wileyonlinelibrary.com]

The task for MC cluster detection represented a signal known exactly, and readers were instructed to look within the center of the ROIs for the clusters. For DBT, observers were instructed to scroll through all slices of the VOIs before finalizing their selection. When displaying the ROIs, the Four-squares program calculated a default window width and level (W/L) based on the range of pixel values within the image. All ROIs had 16-bit unsigned integers. Observers were allowed to adjust the display W/L as needed. The study was performed over multiple sessions, and breaks were encouraged after 25 min of reading to avoid observer fatigue. Prior to the study, reader training was conducted. During this training, the reader was familiarized with the study objective and software interface. The reader scored images from training data independent from the testing data, with supervision from the investigators and feedback provided after each response. During the study, the reader scored training images for each signal, modality, and vendor before scoring the corresponding testing images. Feedback was given after every user response during both training and testing phases. A summary of the number of ROIs scored for each modality is presented in Table III.

Results were computed using the iMRMC[27] package in R Studio (Version 1.1.463). The proportion correct (PC) was calculated as the ratio of correctly selected ROIs to the total number of ROIs scored. In a 4-AFC, the PC for random guessing would be 0.25. The variance of the PC was calculated directly from each trial, and accounts for all correlations across readers and cases. In summary, the variance of PC was computed using u-statistics in the iMRMC package. To do this, the constituent parts of the unbiased variance estimate were first calculated. From these, the statistical moments and their associated coefficients may be derived. Finally, the variance is computed as the inner product of the moments and coefficients. More details of this approach may be found in Gallas et al.[28] Using the variance, the 95% confidence intervals were then calculated as a product of ±1.96 and the standard error. An estimate of detection relative to FFDM as a baseline was computed as ΔPC, defined as $\Delta PC = PC_i - PC_{FFDM}$, where PC_i is the PC of a given modality i (either DBT or SM). Since all ΔPC are relative to FFDM, a value of $\Delta PC > 0$ indicates an improvement in signal detection over FFDM, while a value of $\Delta PC < 0$ indicates a reduction.

TABLE III. Summary of ROIs used in 4AFC study.

Modality	Pixel Size (µm)	Signal	Number of signal present ROIs	Number of signal absent ROIs	ROI Size (pixels)
DBT/ SM	85–105	MC	92–106	276–318	143 x 143 177 x 177
		Mass	44–62	132–186	190 x 190 235 x 235
FFDM	50–100	MC	99–102	297–306	150 x 150 231 x 231
		Mass	44–62	132–186	200 x 200 308 x 308

To determine if differences in PC were statistically significant, P-values and significance level α were required. The P-value was derived via t-table and calculation of the test statistic, computed for every pairwise comparison: signal, vendor, and modality. Then, the P-value can be compared with a Bonferonni-corrected α to determine significance. Computation of the P-values via t-table required an estimate of the number of degrees of freedom (df). The df was estimated as the number of readers, under the assumption that the readers would contribute most to the variability in the results. In addition, having fewer df yields a more conservative estimate of P-values, reducing the likelihood of Type I errors. While a threshold value of $\alpha = 0.05$ is typically used to reject the null hypothesis, the Bonferonni correction is needed in order to account for the increased likelihood of finding statistical significance when there are multiple experiments. In MRMC study design, multiple experiments can arise from comparing the PC across different modalities, vendors, or signals. Thus, having multiple comparisons may require determination of a new threshold for significance α/m, where m is the number of *independent* comparisons. In this study, there were three pairwise comparisons per signal for each vendor with DBT, FFDM, and SM (DBT vs FFDM, DBT vs SM, and FFDM vs SM), resulting in a threshold of $\alpha = 0.05/3 = 0.0166$. The GE Essential (SenoClaire) system did not have SM. Only one pairwise comparison was made (DBT vs FFDM), so the threshold remained $\alpha = 0.05$.

3. RESULTS

Sample images are presented in Fig. 4 for the masses; side-by-side comparisons are given for DBT, FFDM and SM (unless unavailable) for each of the five systems. Arrows indicate the locations of the masses. Some masses become difficult to detect when going from 3D to 2D, especially comparing DBT to SM. For DBT, the masses appear most conspicuous for the systems with the largest gantry spans, namely Siemens at 50°, GE Essential at 25°, and GE Pristina also at 25°.

The reader scores are presented in Table IV for all systems, modalities, and signals. The values displayed are the reader-averaged PC with the 95% confidence interval (CI_{95}) in brackets. The reader-averaged scores are presented for masses in Fig. 5 and for MCs in Fig. 6. The error bars represent one standard deviation accounting for all sources of variability (reader and case). Results are given for DBT with the red bars, for FFDM with green, and for SM with blue. The gantry span is indicated in degrees above each subplot, and the average glandular dose is given below each bar. The pairwise comparisons with an asterisk denote statistical significance with the Bonferroni correction.

For masses, the highest performance was achieved overall with DBT. Furthermore, a relationship was observed between overall PC and gantry span. The PC increased from 0.72 ± 0.05 for Hologic with 15° (and a similar score for

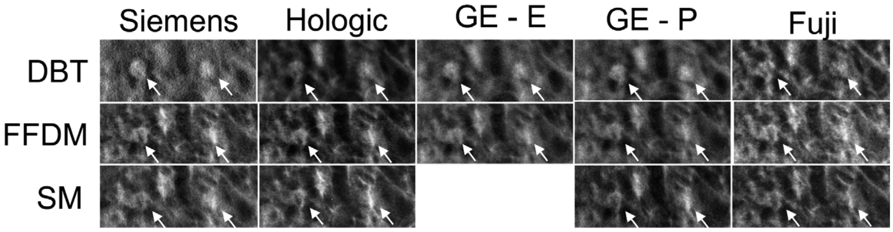

FIG 4. Example of masses within phantom. Regions containing masses are presented for each vendor. The same location was selected in each image with arrows indicating the locations of signals.

Fuji) to 0.91 ± 0.03 for Siemens with 50°. The difference in performance between DBT and FFDM was found to be statistically significant for all systems except Fuji and Hologic — for GE Essential (SenoClaire) the difference between DBT and FFDM had a *P*-value of 0.05, right on the threshold of $\alpha = 0.05$ as the Bonferroni correction was not necessary. The difference between DBT and SM was statistically significant for all systems except Fuji. The difference in performance between FFDM and SM was not found to be statistically significant. For FFDM, the PC for mass detection varied minimally across systems and different dose levels ranging from 0.61 ± 0.06 (Fuji) to 0.64 ± 0.04 (Siemens). Comparably, for SM the PC ranged from 0.52 ± 0.05 (Hologic) to 0.65 ± 0.05 (Fuji). Although the SM image is typically produced using the DBT dataset, no trend was observed between the scores for SM and the gantry span of the system.

For MCs, the highest PCs were observed with FFDM and DBT, both having similar scores. Overall scores for MCs were observed to be higher than those of the masses; with DBT, the PC ranged from 0.84 ± 0.02 (Siemens) to 0.95 ± 0.02 (GE Pristina), while FFDM ranged from 0.78 ± 0.03 (Siemens) to 0.94 ± 0.02 (Hologic). Performance with SM was lowest, with PC ranging from 0.39 ± 0.04 (Siemens) to 0.63 ± 0.05 (Fuji).

TABLE IV. PC scores for all systems with 95% confidence interval (CI_{95}) in brackets.

	Masses		Microcalcifications	
	PC	CI_{95}	PC	CI_{95}
Hologic				
DBT	0.72	[0.62,0.81]	0.93	[0.89,0.97]
FFDM	0.61	[0.51,0.70]	0.94	[0.91,0.97]
SM	0.52	[0.42,0.61]	0.61	[0.54,0.68]
Fuji				
DBT	0.73	[0.64,0.82]	0.87	[0.80,0.93]
FFDM	0.61	[0.50,0.72]	0.84	[0.77,0.91]
SM	0.65	[0.55,0.75]	0.63	[0.54,0.72]
GE Essential				
DBT	0.80	[0.71,0.89]	0.84	[0.79,0.90]
FFDM	0.64	[0.54,0.75]	0.79	[0.73,0.85]
GE Pristina				
DBT	0.84	[0.76,0.91]	0.95	[0.92,0.98]
FFDM	0.62	[0.54,0.70]	0.92	[0.88,0.96]
SM	0.60	[0.50,0.70]	0.53	[0.44,0.62]
Siemens				
DBT	0.91	[0.84,0.97]	0.84	[0.79,0.88]
FFDM	0.64	[0.56,0.71]	0.78	[0.72,0.83]
SM	0.56	[0.46,0.65]	0.39	[0.31,0.46]

The ΔPC relative to FFDM is provided in Fig. 7 for all vendors. Results are given for both masses and MCs side-by-side, with DBT in red and SM in teal. For masses, DBT consistently yielded a positive $\Delta PC > 0.10$. This indicated that detection of masses was higher with DBT than with FFDM regardless of system configuration, for the present task. For MCs, however, moderate improvement was observed with DBT relative to FFDM, with all $\Delta PC \leq 0.06$. Conversely, SM yielded negative ΔPC in all but one comparison, for both mass and MC detection. Moreover, the greatest difference was observed for MCs, indicating that for this size of MCs worse performance will be obtained with SM compared to FFDM.

4. DISCUSSION

The commercial systems investigated in this study varied greatly in design and how they operate. Differences in x-ray spectra, detector type (direct vs indirect-conversion), detector pixel and reconstructed voxel size, acquisition geometry, reconstruction method, image postprocessing methods, and step-and-shoot vs continuous gantry motion could have affected performance depending on the task. In addition, the GE SenoClaire and Pristina systems utilize antiscatter grids while acquiring DBT projections, while the other systems do not. Using acquisition parameters determined from the AEC software of each system, the estimated average glandular dose (AGD) varied between systems, sometimes substantially. Owing to the many system parameters affecting image quality, it is difficult to determine which factors affect performance the most. Therefore, it is difficult to make conclusions on how specific design and acquisition parameters affect the results here and the current study should not be considered a vendor comparison. Nevertheless, certain general trends can be observed in this study. Furthermore, we believe that the resulting phantom images and computed task performance demonstrate that this methodology can be utilized on all clinically available DBT systems.

For the mass detection study, task performance was higher with DBT than with FFDM or SM, and the difference in PC was statistically significant for two systems. This finding concurs with other phantom studies using structured

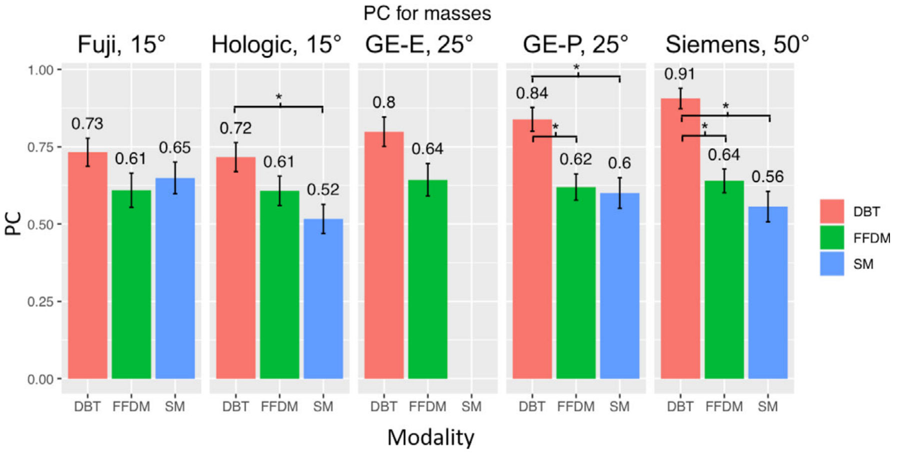

FIG 5. Reader averaged PC for masses. PC was highest with Digital breast tomosynthesis (DBT) across all systems, while full-field digital mammography (FFDM) and SM had similar, lowers PC scores. Asterisks indicate a statistically significant difference. Portions of figure reprinted with permission from Ikejimba et al. "Assessment of task-based performance from five clinical DBT systems using an anthropomorphic breast phantom," 15th International Workshop on Breast Imaging (IWBI2020). Vol. 11513 (2020) Red — "DBT". Green — "FFDM". Blue — "SM". [Color figure can be viewed at wileyonlinelibrary.com]

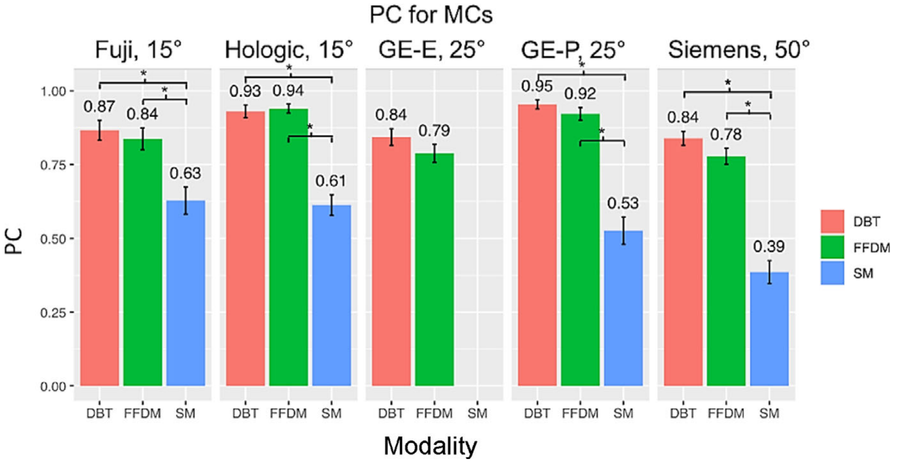

FIG 6. Reader averaged PC for MCs. PC was highest with digital breast tomosynthesis (DBT) and full-field digital mammography across all systems. Asterisks indicate a statistically significant difference. Red — "DBT". Green — "FFDM". Blue — "SM". [Color figure can be viewed at wileyonlinelibrary.com]

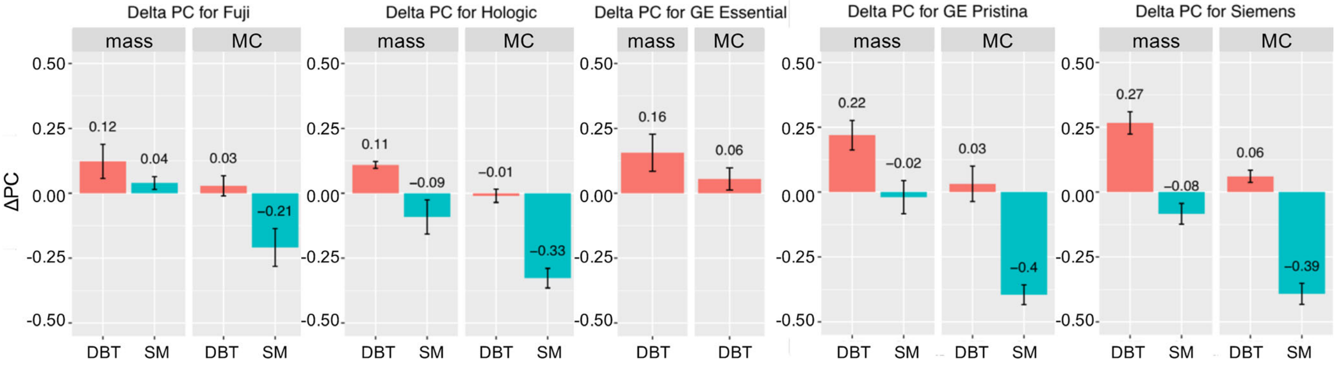

FIG 7. ΔPC for all systems. The ΔPC was computed relative to full-field digital mammography for digital breast tomosynthesis (DBT) and SM, with results given for masses and MCs. [Color figure can be viewed at wileyonlinelibrary.com]

backgrounds. For example, Cockmartin et al.[15] showed clear improvement of DBT over FFDM in detecting masses of various sizes. From a subjective visual impression comparing mass lesions with DBT and FFDM, we concluded that mass conspicuity is generally improved with DBT over all systems (see Fig. 4), which likely relates to the improved diagnostic performance of DBT reported in many lab based clinical studies.[29] Additionally, mass detection performance of the DBT systems trended with increased gantry span, with lowest PC from Hologic and Fuji (both 15°), then both GE systems (25°), and highest for Siemens (50°). A subjective visual impression clearly shows that the wider angle DBT systems provide improved mass lesion conspicuity. This finding concurs with other phantom studies that show a strong correlation between mass visualization and increased gantry angle.[30–35] In particular, previous reserachers[31–33] examined the relationship between mass detectability and DBT gantry span from 15° or 16° up to 60°, along with a number of other acquisition variables. Within the context of that work, the results in the present paper align with the trends observed when matched with a similar gantry span and number of projections. Similar reader scores for the mass detection task were observed between FFDM and SM, a finding that was in agreement with other studies. For example, Mackenzie et al.[36] conducted a virtual clinical trial that showed similar mass detection performance with SM and FFDM. Although there are many variations of SM algorithms implemented by different vendors, clinical studies also seem to suggest similar performance of SM and FFDM for the detection and diagnosis of mass lesions.[4,37,38] For FFDM, the performance of mass detection did not appear to be impacted by dose with PCs of 0.61, 0.64, 0.64, 0.62, and 0.61 measured for reference AGD values of 0.84, 0.90, 0.93, 1.09, and 1.63 mGy, respectively. This finding was in agreement with previously published detection studies using hybrid clinical data, that is, normal patient data inserted with simulated lesions, by Svahn et al.[39] and Timberg et al.[40]

For the MC detection task, performance was similar with DBT and FFDM, with both modalities providing improved performance over SM. Unlike results with mass detection, no clear trend was determined between reader performance and the system geometry. Chan et al.[41] observed a trend of decreasing MC detection sensitivity and conspicuity with increasing scan angle for acquisition with uniform angular increments; however, the DBTs at all scan angles were acquired with a step-and-shoot system, the same x-ray spectrum, dose, and detector in that study. The differences in the many factors among the DBT systems for the current study may have reduced the dependence of MC detection on the scan angle. Due to their size, detection of MCs are probably more limited by quantum noise, whereas the detection of mass lesions are probably more limited by overlapping structure, thus explaining the greater improvement of DBT over FFDM for the mass detection task. This observation was discussed in detail by Burgess et al.[42] showing that smaller microcalcification-like objects have different contrast-detail characteristics than larger mass-like objects. Of course, it is difficult to be certain of this trend because other factors such as detector type and pixel size might also contribute to differences in performance. Unlike the mass detection task, performance of DBT and FFDM with MCs was significantly higher even with the Bonferroni correction for most of the systems tested compared to SM. This result concurs with the findings of Mackenzie et al.[43] who showed that detection of subtle MCs was significantly reduced with SM as compared to DBT and FFDM using simulated lesions inserted into clinical data. Most clinical studies to date have shown that detection of MCs is comparable between SM and FFDM, both alone and with DBT.[37,44,45] However, it is important to note that often these studies include a range of MC sizes present in clinical patient data. In the present study, the sizes of the MCs were restricted to a range of 150–180 μm to interrogate performance with a challenging task. Results showed that detection of MCs were inferior for SM, the lowest PC scores for this size range. While some clinicians have indicated preference for SM when viewing MCs, maybe due in part to over-enhancement for larger MCs, it is possible that the smallest, more subtle MCs may not have been conspicuous on SM.

In Figs. 5 and 6, statistically significant differences are indicated with denoted asterisks. To account for multiple comparisons, Bonferroni correction was used. Increasing the number of comparisons m yields a lower threshold for significance α, representing a more conservative approach to determine statistical significance. While this decreases the chance of Type I errors when rejecting the null hypothesis, it also increases the chance of Type II errors. In this study, a high m value can be justified if comparisons were made across many systems and modalities. However, a value of $m = 2$ or $m = 3$ may be appropriate since the comparisons were mainly made between 2 or 3 modalities from a single manufacturer. In practice, the number of independent comparisons can be difficult to determine, since the same object is imaged across the systems and the same readers are used for assessing the images. It is important to use proper judgment when applying the correction.

The present study examined mass and MC detection viewed by each modality alone. In clinical practice, the current standard of care for breast cancer screening in the US is to observe either a conventional 2D FFDM study alone, a combination of a 2D plus 3D DBT study, or a 3D DBT study alone in the case of Siemens. However, there is a growing trend towards screening with a single modality to minimize radiation exposure to patients and to reduce reading time for exams. For this reason, it is of interest to evaluate detectability of a single modality. The results reported herein suggest that subtle MCs could be missed if reading SM images alone without also reading DBT images.

In this study, systems were compared at the AEC dose levels for each machine, rather than at a fixed dose across the systems. This was done for two reasons. First, according to each vendor, the AEC setting represents the optimal beam conditions for imaging of a specific breast. That is, the AEC parameters are optimized to achieve a certain image quality on a given system. Second, operating at an *a priori* fixed dose could result in an advantage or disadvantage for the system, depending on if the AEC dose is respectively lower or higher than the selected dose. In this study, the reference AEC doses ranged from 0.89 to 1.86 mGy for DBT and from 0.84 to 1.63 mGy for FFDM, representing almost a twofold increase from the lowest to the highest dose. If the dose was fixed to an arbitrary value, it is possible that results would change in a way that was different for each system. This could yield scores of reader performance that may not be reliable, particularly for dose-limited tasks such as MC detection. For these reasons, the imaging data was collected under the standard AEC conditions for each system.

The methodology presented can be a useful tool for routine QC, system optimization, or comparing task-based performance between different imaging systems.[46] Currently, a widely accepted approach for QC involves using the ACR mammography accreditation phantom or similar phantoms.[47,48] The ACR phantom requires subjective reading of signals by human observers (i.e., the reader knows beforehand that the signals are present, and they are asked to record whether the signal is visualized). Less subjective, automated methods for reading the ACR phantom are available;[49–51] however, they are not typically used. CNR measurements can also be used for QC, but CNR does not account for pixel size or task. If model observers are developed, they may be utilized in the present methodology for a fast, task-based quantitative approach to QC. For QC purposes the parchment phantom can be designed to incorporate various known signal types for detection (e.g., fibers, specks, and so on), uniform regions for noise power spectra, additional BBs for point spread functions and azimuthal spread functions, and other features for automatic analysis. This methodology can lend itself to system optimization whereby optimal acquisition or reconstruction parameters may be determined based on mass or MC detectability. This has the advantage that the results are task-specific. To improve with the imaging workflow for optimization studies, the current design of the support plate can be modified to push the posts towards the chest wall, removing the need for separate shots with the posts extending from the detector edge. Lastly, this methodology has promise in regulatory applications for potentially expediting the review process.

While this study was large in its scope, there were a few limitations. First, the phantom modeled a single patient anatomy. A greater range of background anatomy was simulated by placing the signals into different regions of breast parenchyma, but a future study could involve the use of multiple breast phantoms to increase background variability even more. This would then yield a higher number of ROIs for both MCs and masses. As previously mentioned, the phantom modeled a dense breast, so it is possible that overall system performance may differ for a fatty breast more typical of the general population. In addition, the phantom used here was not strictly based on patient data, although it is visually similar. However, it is unclear if phantom realism would impact reader performance differently on different modalities. Another limitation is that the signals represented only one size and shape for the masses and one size range for the MCs, and the print density of the masses was adjusted to make the task challenging. Future studies could investigate performance with both benign- and malignant-appearing lesions, match KI attenuation with known tissue attenuation in the mass, and fabricate MCs comprising different materials[52] other than calcium hydroxyapatite.[53] Additionally, all the MCs were contained within one slice. It is not clear if this would benefit a particular system because of differences in axial resolution. Another potential limitation is that the ROIs were displayed at a 1-to-1 magnification. Because the systems have different voxel sizes, this resulted in the ROIs being displayed at different physical sizes. Consequently, it is possible that displayed ROIs could be smaller than what a radiologist would use if she did not employ additional image magnification. Finally, the readers were not radiologists. Using non-radiologist readers may not have affected the scores since the detection tasks were relatively simple.[54] Nonetheless, human readers can suffer from observer fatigue, and the presence of different skill levels can cause intra- and inter-observer variability. More experienced readers were also observed to perform better than less-experienced readers for some tasks. To circumvent variability due to readers, model observers are being developed for these types of tasks and can be used to minimize variance and reduce the reading time.[55] Still, challenges associated with the wide scale use of phantoms of this type include: reproducibility of the phantom printing process, similarity of multiple phantoms, long-term use of ink with high salt concentration on desktop inkjet printers, and reproducibility across multiple printer brands.

5. CONCLUSION

In this work, we demonstrated the use of an anthropomorphic breast phantom to objectively assess task-based performance of different commercial breast imaging systems. The phantom was imaged on five commercially available DBT systems across four states, and scans were collected with masses and MCs inserted. A 4AFC observer study was conducted to assess performance with FFDM, DBT, and SM. For masses, overall detection was highest using DBT, with an improvement observed with increased gantry span. For MCs, performance was highest with DBT and FFDM and worse with SM. This study is the first of its kind to use a physical inkjet-printed anthropomorphic phantom to assess clinical performance of all commercially available breast imaging systems.

ACKNOWLEDGMENTS

The authors acknowledge the help of Dr. Guo Zhang with the Foursquare software and the help of Andrea Kim with lesion model visualization.

FUNDING

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.

CONFLICT OF INTEREST

H.-P. C. and M.G. have research collaboration with GE through an institutional grant not related to the current study. All other authors have no conflicts of interest to disclose.

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

REFERENCES

1. DeSantis CE, Ma J, Goding Sauer A, Newman LA, Jemal A. Breast cancer statistics, 2017, racial disparity in mortality by state. *CA Cancer J Clin*. 2017;67:439–448.
2. Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Ann Intern Med*. 2003;138:168–175.
3. Mammography Quanlity Standards Act (MQSA) National Statistics; 2020. https://www.fda.gov/radiation-emitting-products/mqsa-insights/mqsa-national-statistics.
4. Garayoa J, Chevalier M, Castillo M, et al. Diagnostic value of the standalone synthetic image in digital breast tomosynthesis examinations. *Eur Radiol*. 2018;28:565–572.
5. Haas BM, Kalra V, Geisel J, Raghu M, Durand M, Philpotts LE. Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening. *Radiology*. 2013;269:694–700.
6. 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.
7. 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.
8. Badano A, Graff CG, Badal A, et al. Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial. *JAMA Netw Open*. 2018;1:e185474.
9. Bakic PR, Barufaldi B, Higginbotham D, et al. Virtual clinical trial of lesion detection in digital mammography and digital breast tomosynthesis. In: *Proc. SPIE 10573, Med. Imaging*. Phys. Med. Imaging; 2018: 1057306.
10. 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.
11. American College of Radiology. *Mammography Quality Control Manual, Medical physicist's section*; 1999:225–330.
12. Bijkerk K, Lindeijer J, Thijssen M. The CDMAM-Phantom: a contrast-detail phantom specifically for mammography. *Radiology*. 1993;185:395.
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. Carton AK, Bakic P, Ullberg C, Derand H, Maidment AD. Development of a physical 3D anthropomorphic breast phantom. *Med Phys*. 2011;38:891–896.
15. 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.
16. 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.
17. Ikejimba LC, Salad J, Graff CG, et al. Assessment of task-based performance from five clinical DBT systems using an anthropomorphic breast phantom. In: *Proc. SPIE Vol. 11513, 15th Int Workshop on Breast Imaging* SPIE; 2020:1151305.
18. Personal communication with Andy Smith of Hologic Inc.
19. Ikejimba LC, Salad J, Graff CG, et al. 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. *Med Phys*. 2019;46:3883–3892.
20. Graff CG. A new open-source multi-modality digital breast phantom. In: *Proc. SPIE 9783, Med. Imaging*. Phys. Med. Imaging; 2016:978309.
21. de Sisternes L, Brankov JG, Zysk AM, Schmidt RA, Nishikawa RM, Wernick MN. A computational model to generate simulated three-dimensional breast masses. *Med Phys*. 2015;42:1098–1118.
22. IEC. 61223-3-6 EVALUATION AND ROUTINE TESTING IN MEDICAL IMAGING DEPARTMENTS - Parts 3-6: Acceptance and constancy tests of mammographic X-ray equipment used in a mammographic tomosynthesis mode of operation.
23. Dance D, Skinner C, Young K, Beckett J, Kotre C. Additional factors for the estimation of mean glandular breast dose using the UK mammography dosimetry protocol. *Phys Med Biol*. 2000;45:3225.
24. Dance D, Young K, Van Engen R. Further factors for the estimation of mean glandular dose using the United Kingdom, European and IAEA breast dosimetry protocols. *Phys Med Biol*. 2009;54:4361.
25. Dance D, Young K, Van Engen R. Estimation of mean glandular dose for breast tomosynthesis: factors for use with the UK, European and IAEA breast dosimetry protocols. *Phys Med Biol*. 2010;56:453.
26. Zhang G, Cockmartin L, Bosmans H. A four-alternative forced choice (4AFC) software for observer performance evaluation in radiology. *Proc SPIE*; 2016;9787.
27. Obuchowski NA, Gallas BD, Hillis SL. Multi-reader ROC studies with split-plot designs: a comparison of statistical methods. *Acad Radiol*. 2012;19:1508–1517.
28. Gallas BD, Bandos A, Samuelson FW, Wagner RF. A framework for random-effects ROC analysis: biases with the bootstrap and other variance estimators. *Commun Stat Theory Methods*. 2009;38:2586–2603.
29. Vedantham S, Karellas A, Vijayaraghavan GR, Kopans DB. Digital breast tomosynthesis: state of the art. *Radiology*. 2015;277:663–684.
30. Goodsitt MM, Chan H-P, Schmitz A, et al. Digital breast tomosynthesis: studies of the effects of acquisition geometry on contrast-to-noise ratio and observer preference of low-contrast objects in breast phantom images. *Phys Med Biol*. 2014;59:5883.
31. Gang GJ, Lee J, Stayman JW, et al. Analysis of Fourier-domain task-based detectability index in tomosynthesis and cone-beam CT in relation to human observer performance. *Med Phys*. 2011;38:1754–1768.
32. Sechopoulos I, Ghetti C. Optimization of the acquisition geometry in digital tomosynthesis of the breast. *Med Phys*. 2009;36:1199–1207.
33. Reiser I, Nishikawa R. Task-based assessment of breast tomosynthesis: effect of acquisition parameters and quantum noise a. *Med Phys*. 2010;37:1591–1600.
34. Samei E, Thompson J, Richard S, Bowsher J. A case for wide-angle breast tomosynthesis. *Acad Radiol*. 2015;22:860–869.
35. Scaduto DA, Huang H, Liu C, et al. Impact of angular range of digital breast tomosynthesis on mass detection in dense breasts. In: Proc. Vol. 10718, 14th Int Workshop on Breast Imaging (IWBI 2018). SPIE; 2018:107181V.
36. Mackenzie A, Kaur S, Elangovan P, Dance D, Young K. Comparison of synthetic 2D images with planar and tomosynthesis imaging of the breast using a virtual clinical trial. In: Proc. SPIE 10577, Med. Imaging. Img Perception, Obsv Perf, and Tech Assessment; 2018:105770H.
37. Choi JS, Han B-K, Ko EY, et al. Comparison between two-dimensional synthetic mammography reconstructed from digital breast tomosynthesis and full-field digital mammography for the detection of T1 breast cancer. *Eur Radiol*. 2016;26:2538–2546.
38. Zuley ML, Guo B, Catullo VJ, et al. Comparison of two-dimensional synthesized mammograms versus original digital mammograms alone

and in combination with tomosynthesis images. *Radiology.* 2014;271:664–671.

39. Svahn T, Hemdal B, Ruschin M, et al. Dose reduction and its influence on diagnostic accuracy and radiation risk in digital mammography: an observer performance study using an anthropomorphic breast phantom. *Br J Radiol*. 2007;80:557–562.
40. Timberg P, Ruschin M, Båth M, et al. Potential for lower absorbed dose in digital mammography: A JAFROC experiment using clinical hybrid images with simulated dose reduction. Proc SPIE. 2006;6146:614614.
41. 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(3):675–685.
42. Burgess AE, Jacobson FL, Judy PF. Human observer detection experiments with mammograms and power-law noise. *Med Phys*. 2001;28:419–437.
43. Mackenzie A, Thomson E, Elangovan P, et al.An observer study to assess the detection of calcification clusters using 2D mammography, digital breast tomosynthesis, and synthetic 2D imaging. In: Proc. SPIE 01952, Med. Imaging. Img Perception, Obsv Perf, and Tech Assessment; 2019:109520U.
44. Choi JS, Han B-K, Ko EY, Kim GR, Ko ES, Park KW. Comparison of synthetic and digital mammography with digital breast tomosynthesis or alone for the detection and classification of microcalcifications. *Eur Radiol*. 2019;29:319–329.
45. Lai Y-C, Ray KM, Lee AY, et al. Microcalcifications detected at screening mammography: synthetic mammography and digital breast tomosynthesis versus digital mammography. *Radiology.* 2018;289:630–638.
46. Makeev A, Ikejimba LC, Salad J, Glick SJ. Objective assessment of task performance: a comparison of two FFDM detectors using an anthropomorphic breast phantom. *J Med Image*. 2019;6:043503.
47. VOXMAM phantom. Leeds Test Objects. https://www.leedstestobjects.com/index.php/phantom/voxmam-phantom/
48. TOR MAM.Leeds Test Objects. https://www.leedstestobjects.com/index.php/phantom/tor-mam/
49. Chakraborty DP, Eckert MP. Quantitative versus subjective evaluation of mammography accreditation phantom images. *Med Phys*. 1995;22:133–143.
50. Gagne RM, Gallas BD, Myers KJ. Toward objective and quantitative evaluation of imaging systems using images of phantoms. *Med Phys*. 2006;33:83–95.
51. Gennaro G, Ferro F, Contento G, Fornasin F, Di Maggio C. Automated analysis of phantom images for the evaluation of long-term reproducibility in digital mammography. *Phys Med Biol*. 2007;52:1387.
52. 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.
53. Makeev A, Ghammraoui B, Badal A, Graff CG, Glick SJ. Classification of breast calcifications in dual-energy FFDM using a convolutional neural network: simulation study. In: Proc. SPIE 11312, Med. Imaging. Phys. Med. Imaging; 2020:113120M.
54. Elangovan P, Mackenzie A, Dance DR, Young KC, Wells K. Using non-specialist observers in 4AFC human observer studies. In: Proc. SPIE 10132, Med. Imaging. Phys. Med. Imaging; 2017:1013256.
55. Petrov D, Marshall N, Young K, Bosmans H. Model and human observer reproducibility for detecting microcalcifications in digital breast tomosynthesis images. In: Proc. SPIE 10577, Med. Imaging. Img Perception, Obsv Perf, and Tech Assessment; 2018:105770B.