Title

Deep Learning to Simulate Contrast-enhanced Breast MRI of Invasive Breast Cancer.

Permalink

https://escholarship.org/uc/item/7f54n11d

Journal

Radiology, 306(3)

Authors

Chung, Maggie
Calabrese, Evan
Mongan, John
et al.

Publication Date

2023-03-01

DOI

10.1148/radiol.213199

Peer reviewed

Feasibility of Simulated Contrast-enhanced Breast MRI for Imaging Malignant Masses Using Deep Learning

Maggie Chung, MD[a,§], **Evan Calabrese, MD, PhD**[a,§], **John Mongan, MD, PhD**[a], **Kimberly M. Ray, MD**[a], **Jessica H. Hayward, MD**[a], **Tatiana Kelil, MD**[a], **Ryan Sieberg, MD**[a], **Nola Hylton, PhD**[a], **Bonnie N. Joe, MD, PhD**[a], **Amie Y. Lee, MD**[a]

[a]Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA

Abstract

Background: There is increasing interest in non-contrast breast MRI alternatives for tumor visualization to increase the accessibility of breast MRI.

Purpose: To evaluate the feasibility and accuracy of generating simulated contrast-enhanced T1-weighted breast MRIs from pre-contrast MRIs in biopsy-proven invasive breast cancer using deep learning.

Methods and Materials: Women with invasive breast cancer and contrast-enhanced breast MRI performed for initial evaluation of extent of disease were retrospectively identified between January 2015 and December 2019 at a single academic institution. A three-dimensional, fully convolutional deep neural network simulated contrast-enhanced T1-weighted breast MRIs from five pre-contrast sequences (T1-weighted non-fat-suppressed [FS], T1-weighted FS, T2-weighted FS, apparent diffusion coefficient, and diffusion-weighted imaging). For qualitative assessment, four blinded breast radiologists (3 to 15 years of experience) assessed image quality (excellent/acceptable/good/poor/unacceptable), presence of tumor enhancement, and maximum index mass size using 22 pairs of real and simulated contrast-enhanced MRIs. Quantitative comparison was performed using whole breast similarity and error metrics and Dice coefficient analysis of enhancing tumor overlap.

Results: 96 MRIs from 96 women (mean age, 52 years $\pm$ 12 [SD]) were evaluated. The readers assessed all simulated MRIs as having the appearance of a real MRI with tumor enhancement. Index mass sizes on real and simulated MRIs demonstrated good-to-excellent agreement (intraclass correlation coefficient, 0.73–0.86; *P*<.001) without significant differences (mean differences −0.8–0.8 mm, *P*=.36-.80). Almost all simulated MRIs (84 of 88; 95%) were considered of diagnostic quality (ratings of excellent/acceptable/good). Quantitative analysis demonstrated strong similarity (structural similarity index, 0.88 $\pm$ 0.05), low voxel-wise error (symmetric mean absolute percent error, 3.26%), and Dice coefficient of enhancing tumor overlap of 0.75 $\pm$ 0.25.

Conclusions: It is feasible to generate simulated contrast-enhanced breast MRI using deep learning. Simulated and real contrast-enhanced MRI demonstrated comparable tumor sizes,

Corresponding Author: Maggie Chung, mchung2017@gmail.com, Telephone: 415-476-8358.
§These authors contributed equally in this work

areas of tumor enhancement, and image quality without significant qualitative or quantitative differences.

Summary Statement:

Simulated contrast-enhanced breast MRIs generated using deep learning demonstrate no significant quantitative or qualitative differences compared to real contrast-enhanced MRIs, and have the potential to increase the accessibility of breast MRI.

INTRODUCTION:

Contrast-enhanced breast MRI is the most sensitive test for the detection of breast cancer. There is increasing interest in broadening MRI supplemental screening to women with dense breasts and women at average to intermediate risk (1,2). While gadolinium-based contrast agents (GBCAs) are widely accepted as safe, there are potential limitations associated with their use. The administration of GBCAs requires intravenous access and physician monitoring, which increases the cost and length of breast MRIs. The use of GBCAs also limits MRI in the small subsets of patients with severely impaired renal function and pregnant patients. Furthermore, concerns remain about the unknown clinical significance of gadolinium deposition, especially in the supplemental screening population (3). In response, there is rising interest in non-contrast alternatives in breast MRI. Studies to date have shown that the diagnostic performance of non-contrast sequences (diffusion-weighted imaging in conjunction with T1 and/or T2 series) is promising but remains inferior to contrast-enhanced MRI (4–10).

Recent studies using brain MRI studies have demonstrated the feasibility of synthesizing contrast-enhanced images from pre-contrast inputs using deep learning (11–14). These results raise the possibility of applying similar deep learning techniques to breast MRI. However, the feasibility of creating simulated contrast-enhanced MRI in the breast is unknown. Compared with the brain, the breast poses several unique challenges, including misregistration due to mobility of the breast tissue, greater field inhomogeneity, fewer pre-contrast series, and wider variability in tissue density and vascularity.

The purpose of this pilot study was to evaluate the feasibility and accuracy of generating simulated contrast-enhanced T1-weighted breast MRI from pre-contrast MRI sequences in patients with biopsy-proven invasive breast cancer using deep learning.

METHODS AND MATERIALS:

Our institutional review board approved this Health Insurance Portability and Accountability Act-compliant study and waived the requirement for written informed consent.

Patient Sample

A single academic institution, retrospective radiology database search identified consecutive breast MRI examinations performed between January 2015 and December 2019 for extent of disease evaluation in women with invasive breast cancer. Exclusion criteria were ductal carcinoma in situ without invasive disease; post-neoadjuvant treatment; MRI performed for

evaluation of residual/recurrent disease; and work-up of metastasis with unknown location of breast primary.

MRI Acquisition, Image Preprocessing, Deep Learning Convolutional Neural Network Architecture, and Model Hyperparameters are detailed in Figure 1 and Appendix.

Model Inputs and Training

The network was trained to predict post-contrast images from 5 pre-contrast sequences: T1-non-FS, T1-FS, T2-FS, diffusion-weighted imaging, and apparent diffusion coefficient. The first post-contrast phase of the dynamic post-contrast T1-FS series was used as the ground truth. A seven-fold cross-validation method was used with a split of 85% training and 15% testing, ensuring independent training and testing sets. Full code for network implementation is provided at https://github.com/ecalabr/breast_simulated_gad. Additional details in Appendix.

Quantitative Assessment: Whole Breast Similarity and Error Metrics

Quantitative analysis was performed on the test set including images from all seven cross-validation iterations. Similarity between real and simulated MRIs was quantitatively evaluated using three similarity metrics (neighborhood cross-correlation, histogram mutual information, structural similarity index) and four error metrics (normalized root mean square error, symmetric mean absolute percent error, log accuracy ratio, and median symmetric accuracy) computed across the whole breast (15–17). Additional details in Appendix.

Quantitative Assessment: Dice Overlap of Enhancing Tumor Volume

Enhancing components of biopsy-proven malignant masses were segmented on real and simulated MRI. We excluded studies with only non-mass enhancement due to ill-defined boundaries limiting segmentation reproducibility. Tumor segmentations on real and simulated MRI were compared using Dice coefficient of overlap.

Qualitative Assessment: Multi-reader Study

To supplement the quantitative assessments, a multi-reader study was performed. Thirty pairs of simulated MRI and real MRI from the test sets of two cross-validation folds were obtained for the multi-reader study. A breast radiology attending (A.L.), with six years of experience and who was not a study reader, reviewed the real MRIs while blinded to the simulated MRIs. The radiologist identified 22 cases with an identifiable and measurable mass on the real MRI which were then included in the reader study. Cases without an identifiable index mass on the real MRI and cases with only foci and/or non-mass enhancement were excluded. The 22 pairs of real and simulated MRI were separated, randomly assigned, and read between the two sessions separated by a minimum two-week washout period. Each session included a mix of simulated and real post-contrast images in randomized order. The real and simulated MRI from the same patient never appeared together in the same reader session. The 22 pairs of real and simulated MRI were independently assessed by four blinded breast imaging radiologists (authors B.N.J., K.M.R., T.K., J.H.H.) with experience ranging from 3 years to 15 years. Readers indicated whether the whole exam had the appearance of a real contrast-enhanced breast MRI and

whether the index tumor demonstrated enhancement. Readers measured the maximum axial dimension of the dominant index mass. Readers assessed diagnostic image quality on a 5-point Likert-type scale (5: excellent, 4: good, 3: acceptable, 2: poor, 1: unacceptable). Additional details are in Appendix.

Statistical Analysis

Intraclass correlation coefficients quantified the correlation between tumor sizes on real and simulated MRI by reader as well as inter-reader reliability among the four readers. Modified Bland-Altman plots with repeated measures assessed tumor size agreement between simulated and real MRI (18). Mean bias values and limits of agreement lines (two standard deviations above and below the mean) were calculated. Paired t-test evaluated differences between the tumor sizes on real and simulated MRI. Quality ratings from the four readers were summed to obtain a combined rating for real and simulated MRI, respectively. The number of studies considered acceptable for diagnostic use (quality score 3) were compared using McNemar's test. Similarity and error metrics between the full model and each combination of T1 FS plus one pre-contrast series were compared using Wilcoxon signed-rank test. P value < 0.05 was considered statistically significant. Statistical analyses were performed by two of the authors (M.C. and R.S.).

Post-hoc analysis of the multi-reader study (n=22) demonstrated a power of 0.99 to detect a clinically significant difference in tumor size (>5 mm) between real and simulated MRI. This was conservatively calculated using the smallest effect size among the four readers (effect size 0.90 based on standard deviation of difference of 5.53) and 2-sided matched pair t-test at the 5% significance level. The number of cases in the multi-reader study is similar to that in the multi-reader study (n = 20) reported by Kleesiek et al. for simulated contrast in brain MRI (12). Statistical analysis was performed using R (version 4.1.2; R Foundation for Statistical Computing) (19). Post-hoc analysis was performed using G*power (20).

RESULTS

Patient Characteristics

101 studies from 101 women were included for training and testing. In each cross-validation fold, the training and testing sets consisted of approximately 86 MRIs and 15 MRIs, respectively. No validation set was included. 101 pairs of real and simulated MRIs were compiled from test sets of the seven cross-validation folds. After training and testing, five more cases were subsequently found to meet exclusion criteria upon additional review of the electronic medical records (not apparent from the initial radiology report review), and these were excluded from quantitative and qualitative assessments. One case was downgraded from invasive lobular carcinoma to lobular carcinoma in situ upon a second review of outside pathology. One case was performed for evaluation of residual/recurrent disease in a patient with history of partial mastectomy. Three cases had pathologic diagnoses of ductal carcinoma in-situ without evidence of invasive ductal carcinoma. 96 women (mean age, 52 years ± 12 [standard deviation]; median, 50 years; interquartile range, 43–60 years) were included for assessment (Figure 2). MRI appearance of the 96 biopsy-proven invasive breast cancers included 51 masses (53%), 8 non-mass enhancement (8%), and 37 mass

and non-mass enhancement (39%). Invasive cancer types were 78 ductal (81%), 10 lobular (10%), 6 ductal and lobular (6%), and 2 mucinous (2%). Mean lesion size was 2.4 cm (median 2.0 cm, range 0.7 – 7.8 cm). Most (71/96; 74%) demonstrated type 3 enhancement kinetics. Demographic information and tumor characteristics are summarized in Tables 1 and 2.

Quantitative Assessment: Whole Breast Similarity and Error Metrics

96 pairs of real and simulated MRIs were included in the assessment of similarity and error metrics across the whole breast (Figure 3). There was strong similarity between real and simulated MRIs in terms of structural similarity index (mean 0.88 ± 0.05 [SD]; median 0.89; IQR 0.86–0.91) and histogram mutual information (mean 0.68 ± 0.15; median 0.70; IQR 0.62–0.77). Neighborhood cross-correlation was lower (mean 0.50 ± 0.11; median 0.52; IQR 0.46–0.57). There was low voxel-wise error between real and simulated MRI as demonstrated by the symmetric mean absolute percent error (mean $3.26\% \pm 1.0\%$; median 3.07%; IQR 2.69–3.47%), normalized root mean square error (mean 0.04 ± 0.01; median 0.04; IQR 0.03–0.04), log accuracy ratio (mean 0.07 ± 0.02; median 0.06; IQR 0.05–0.07), and median symmetric accuracy (mean 0.05 ± 0.02; median 0.05; IQR 0.04–0.05).

Quantitative Assessment: Tumor Dice Overlap

Eight cases with only non-mass enhancement without a discrete mass were excluded from Dice quantitative assessment. After exclusions, 88 pairs of real and simulated MRIs were included in the Dice assessment. Average Dice coefficient of overlap between the real and simulated MRI enhancing tumor component was 0.75 ± 0.25 with median of 0.84 and IQR of 0.75–0.89. Seven cases that failed to simulate enhancement of the index lesion on simulated MRIs were included in the assessment. Additional details are in the Appendix.

Qualitative Assessment: Multi-reader Study

Each of the twenty-two pairs of real and simulated MRIs was read by four blinded radiologists for a total of 176 reads (88 real and 88 simulated MRI reads). All cases of real and simulated MRIs were assessed by the readers as having the appearance of a real contrast-enhanced MRI (Figure 4). One case that failed to simulate enhancement on simulated MRI was excluded from measurement assessment (Figure 5).

Bias of simulated MRI, as assessed using the mean differences between tumor sizes on real and simulated MRI for each of the four readers, were −0.8 mm (95% CI: −3.1 mm, 1.6 mm), 0.8mm (95% CI: −1.1, 2.8 mm), −0.5 mm (95% CI: −2.5 mm, 1.6 mm), and −0.3 mm (95% CI: −2.2 mm, 2.8 mm), respectively (P=.36-.80). Means of the absolute difference within pairs of real and simulated MRI were 3.0 mm (95% CI: 1.9 mm, 4.9 mm), 3.0 mm (95% CI: 1.6 mm, 4.3 mm), 3.1 mm (95% CI: 1.6 mm, 4.6 mm), and 3.2 mm (95% CI: 1.2 mm, 5.2 mm). 85% (71/84) of the tumor sizes on real and simulated MRI demonstrated absolute differences   5 mm. There was good to excellent correlation between real and simulated MRI tumor sizes (intraclass correlation coefficient [ICC] 0.73–0.86; P=<.001). Bland–Altman analyses of tumor size agreement demonstrated mean bias of −0.2 mm (Figure 6). There was good inter-reader reliability among the four readers with ICC of

0.78 (95% CI: 0.52, 0.90)(P=<.001) using real MRI and 0.85 (95% CI: 0.67, 0.94)(P=<.001) using simulated MRI.

Most of the real MRIs (47/88; 53%) were rated as "excellent," and most simulated MRIs (38/96; 43%) were rated as "good." The four readers considered 88/88 (100%) of the real MRIs and 84/88 (95%) of simulated MRIs of diagnostic quality (image quality scores of acceptable/good/excellent) (P=.13). Of the four simulated MRIs rated non-diagnostic, two were due to image resolution, one due to image resolution and incomplete fat saturation, and one due to present but decreased contrast enhancement.

DISCUSSION:

There is increasing interest in the accessibility of non-contrast breast MRI alternatives for tumor visualization. Our results demonstrate the feasibility of using deep learning to generate simulated contrast-enhanced breast MRIs from pre-contrast images. Real and simulated MRIs were quantitatively similar with high structural similarity index (0.88 ± 0.05), low pixelwise error (symmetric mean absolute percent error of 3.26%), and high degree of enhancing tumor overlap (Dice coefficient, 0.75 ± 0.25). There was good to excellent correlation between real and simulated MRI tumor sizes (intraclass correlation coefficient 0.73–0.86; P<.001). Also, in our multi-reader study, 95% (84/88) of simulated MRI had image quality scores of acceptable, good, or excellent and all simulated MRIs were assessed as having the appearance of a real contrast-enhanced MRI.

Two previously published studies used deep learning techniques to generate simulated contrast-enhanced brain MRIs in humans (12,14). But application to breast MRI has several additional challenges. The breast is mobile and compressible with varying background parenchymal enhancement and respiratory motion artifacts, which could affect algorithm performance. Standard breast MRI protocols consist of fewer pre-contrast sequences than brain MRI protocols for post-contrast prediction and lack non-contrast perfusion techniques such as arterial spin labeling. Despite these challenges, our quantitative analysis demonstrated strong quantitative similarity and low voxel-wise error between real and simulated MRI. Our structural similarity index across the whole breast (0.88) was comparable to indices reported in prior brain MRI studies (0.86–0.87)(12,14). Our high Dice score performance (Dice coefficient, 0.75) in enhancing tumor regions between real and simulated MRI was similar to the performance in previous mouse models and surpasses that in human brain tumors (0.65)(13,14). Also, a sizable percentage (95%) of simulated MRIs in our multi-reader study were considered of diagnostic quality, similar to that reported by Kleesiek et al. (12). Most of simulated MRIs were assessed as "good" (43%) while most of real MRIs were assessed as "excellent" (53%). Future refinement of co-registration among MRI series and expansion of training to larger datasets may help improve image quality. Overall, simulated MRI did not demonstrate substantial over- or under-estimation of enhancing tumor size as 95% CIs of bias were within +/− 5 mm, a range considered to be concordant with the real MRI (21).

Our study has limitations. First, this was a small pilot study limited to a single academic institution, limiting the generalizability. Second, for the purposes of this proof-of-concept

study, we focused on invasive breast cancers only. Third, we excluded cases with only non-mass enhancement in the multi-reader study and assessment of Dice overlap of enhancing tumor. Fourth, we did not evaluate extent of disease or include normal, BI-RADS 1 MRI examinations or benign pathologies to evaluate performance metrics. Fifth, lesion measurements were made only in the axial plane. Sixth, misregistration among series may limit algorithm training and performance. Seventh, laterality and location of the index malignancy were provided to the readers. Finally, given the relatively small number of cases in the reader study, readers may have been susceptible to recall bias despite the two-week washout period.

In conclusion, this proof-of-concept study demonstrated the feasibility of generating simulated contrast-enhanced breast MRI using deep learning. Simulated and real contrast-enhanced MRI demonstrated comparable tumor sizes, areas of tumor enhancement, and image quality without significant qualitative or quantitative differences. Although simulated contrast-enhanced MRI is not intended nor likely to replace all real MRI, it has the potential to extend the benefits and accessibility of breast MRI by reducing the necessity of contrast for tumor visualization. In addition, simulated MRI may have potential roles in high-risk screening, although further research is needed. Additional areas for potential investigation include supplemental screening of average to intermediate-risk women and locoregional staging of pregnancy-associated breast cancers. Also, future studies with larger, external datasets should investigate whether specific patient and tumor characteristics (e.g. size, histologic and molecular subtype) may be associated with failed enhancement and evaluate extent of disease, including more lesion types, (e.g. ductal carcinoma in situ and benign pathologies) and other lesion morphologies, (e.g. non-mass enhancement and foci). In particular, further study is needed to evaluate the performance of simulated contrast-enhanced MRI in smaller lesions and those with progressive or plateau kinetics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

APPENDIX

MRI Acquisition:

Bilateral dynamic contrast-enhanced breast MRI examinations were performed on a 1.5-Tesla scanner (Signa; GE Medical Systems) or 3.0-Tesla scanner (Magnetom Verio, Siemens) with a dedicated breast coil (Sentinelle Medical). Images of the bilateral breasts were acquired in the axial plane with patients in the prone position. Imaging sequences included T1-weighted non-fat-suppressed series (T1-non-FS), T2-weighted fat-suppressed series (T2-FS), diffusion-weighted imaging (DWI), and pre-contrast and dynamic post-contrast T1-weighted fat-suppressed series (T1-FS). Examinations were performed using the following parameters: TR/TE for 1.5T imaging, 9/4.4; TR/TE for 3T images, 7.1/4.9; number of excitations, 1; matrix, $512 \times 320 \times 384$; field of view 29–36 cm; slice thickness 2mm at 1.5T or 0.8mm at 3T. Patients received 0.1 mmol/kg bodyweight of a gadolinium-based contrast agent which was administered intravenously with a remote-

controlled power injector (Spectris; Medrad) followed by 20 mL saline flush. Acquisition times were approximately180 seconds each. The central phase-encoding lines of each dataset were acquired halfway through the acquisition, resulting in an effective sample time of approximately 90 seconds for the early postcontrast series. The diffusion sequences were acquired after dynamic contrast-enhanced breast MRI examinations. Diffusion-weighted imaging data were acquired with a fat-suppressed diffusion-weighted echo-planar imaging sequence using the imaging parameters: TR/TE = 6000/69.6 ms, b = 0, 600 s/mm^2, FOV = 400×400 mm, matrix = 128×128, slice thickness = 3 mm, gap = 0, averages = 6, and voxel size: = 29.3 mm^3.

MRI Pre-processing: Image Data Conversion and Deidentification

Training data consisted of T1-non-FS, T2-FS, pre-contrast and post-contrast T1-FS, and DWI trace. The first post-contrast phase of the dynamic post-contrast T1-FS, obtained at approximately 90 seconds, was the ground-truth. All MRI examinations underwent pre-processing steps including data conversion from DICOM to NifTI, deidentification, co-registration, whole breast segmentation, and intensity normalization. All image data were deidentified and converted to the NIfTI file format using dcm2niix 1.0. The dcm2niix command structure used was the following: dcm2niix -w 2 -b n -z y -x n -t n -m n -f /path/to/output/nifti.nii.gz -o /path/to/output/ -s y -v n /path/to/dicom/image.dcm.

MRI Pre-processing: Co-registration

Co-registration of all images to a standardized anatomic space was performed using freely available automated image registration tools (Advanced Normalization Tools [ANTS])(22). T1-non-FS, T2-FS, and post-contrast T1-FS sequences were registered to the T1-FS using a rigid + affine registration strategy. For DWI, the B = 0 sec/mm^2 images were registered to the T2-fat-suppressed using a rigid + affine registration strategy then applied to the trace DWI images. The T1-FS images were selected as the registration target because these images have the highest spatial resolution and contrast-to-noise ratio of the non-contrast series. Registration parameters were as follows: rigid: gradient step 0.1, four levels with shrink factors of [4, 3, 2, 1], smoothing sigmas of [6, 4, 1, 0] voxels, iterations of [1000, 500, 250, 50], convergence threshold of 1e-07, and a convergence window of 10 samples; affine: gradient step 0.1, four levels with shrink factors of [4, 3, 2, 1], smoothing sigmas of [6, 4, 1, 0] voxels, iterations of [1000, 1000, 1000, 1000], convergence threshold of 1e-07, and a convergence window of 10 samples. In all cases, mutual information was used as the similarity metric with 32 bins, a regular sampling strategy, and a sampling percentage of 25%.

For quality control, each study was manually inspected by one of the authors (M.C.) for accurate co-registration among all series by checking several points through the MRI exam including the breast-air interface, chest wall-to-mediastinum interface, and lymph nodes. Exams with misregistration of greater than 5 mm were excluded. See Supplemental Figure 1 for an example of pre- and post-registration series images.

MRI Pre-processing: Whole Breast Masking

De-identified data underwent a series of pre-processing steps implemented using Python 3.8 and Nipype 1.6.0. Automated whole breast segmentation was performed on T1 non-FS using a 3D deep learning, convolutional neural network segmentation algorithm based on the U-net architecture with 2 downsampling steps and 3, 4, and 5 convolutional layers per level (14,23). We trained the network using overlapping $80 \times 80 \times 80$ voxel patches from 200 breast MRIs that were manually segmented to omit areas including air, mediastinum, and liver from each image using the ITK-SNAP 3.8.0 software. All automated whole breast segmentations were inspected by one of the authors (M.C.). When necessary, manual correction was performed by the author using ITK-SNAP 3.8.0 for identified areas of under-or over-segmentation (24). Automated whole breast segmentations were applied as masks to the other co-registered MRI series.

MRI Pre-processing: Intensity Normalization

Each individual image series underwent intensity normalization during pre-processing using the zero mean, unit standard deviation method (sometimes referred to as Z-score). CNR and SNR were not separately calculated and compared.

Deep Learning, Convolutional Neural Network Architecture

We implemented a previously published three-dimensional deep learning, convolutional neural network model used by Calabrese et al. to synthesize post-contrast images from pre-contrast brain MRI (14). The model architecture was based on the U-net architecture (25) with three-dimensional convolutions, convolutional up/downsampling, long range skip connections, bottleneck residual blocks (26), per-layer batch normalization, leaky ReLU activation, and feature dropout (Figure 1). The U-Net architecture consisted of three levels (including the bottom level) with three, four, and five $3 \times 3 \times 3$–voxel convolution bottleneck residual blocks, respectively. In the descending (encoding) limb of the network, the first two levels were followed by a $1 \times 1 \times 1$– voxel convolutional downsampling layer with strides of $2 \times 2 \times 2$ voxels. In the ascending (decoding) limb of the network, the last two levels were preceded by a $1 \times 1 \times 1$ deconvolutional (transpose convolution) upsampling layer with strides $2 \times 2 \times 2$ voxels. A $1 \times 1 \times 1$ convolutional layer with a single filter was used as the output layer to map features directly to output image intensities. Long-range skip connections between the descending and ascending limbs were accomplished with feature concatenation. The number of filters per layer was determined by multiplying a base number of filters by 2 for each downsampling layer and dividing by 2 for each upsampling layer. The final model had a total of 62 3D convolutional layers and 835,505 trainable parameters. The model was implemented using Python 3.8 and Tensorflow 2.3.0. Full code for network implementation is provided at https://github.com/ecalabr/breast_simulated_gad (specific version hash: 76e5347b5e9b0e98ab02db1b1728986412050de8).

Model Hyperparameters

We used the optimal hyperparameters reported by Calabrese et al. given the similarity of the deep learning tasks (14). Also, the relatively low data volume in this pilot study limited a robust hyperparameter search. The hyperparameters included 32 base filters per layer, batch size of 16, and dropout rate of 0.4. The hyperparameters were not tuned using the training or validation data.

Model Inputs and Training

The network was trained to predict post-contrast images from 4 pre-contrast sequences: T1-non-FS, T1-FS, T2-FS, and DWI. The first post-contrast phase of the dynamic post-contrast T1-FS series, obtained at approximately 90 seconds, was the ground-truth. Training inputs consisted of $80 \times 80 \times 80$–voxel patches from each of the four precontrast image series with strides of 20 voxels in all dimensions (25% overlap). The different series contrasts were concatenated in a 4th dimension (total input data shape of $80 \times 80 \times 80 \times 4$). Patches were uniformly sampled across the masked breast volume. Patches with less than 20% of voxels within the breast (i.e. >80% empty) were discarded. Data augmentation steps included random three-axis rotations of −30 to +30 degrees and random dimension swaps. A seven-fold cross-validation method was used. For each cross-validation fold, we reinitialized a blank model, independently trained it with 85% of MRIs, and used the remaining 15% as independent test data. Validation sets were not included. Training consisted of predetermined 20 epochs without early stopping. The loss function for model training was the mean squared error between the intensity-normalized real and simulated contrast-enhanced images. The test set rotates through the data set such that, after all seven cross-validation folds, all 101 MRIs are included in the test set once. After completing all seven cross-validation folds, we compiled the simulated contrast MRIs (n=101) from the seven test sets for qualitative and quantitative analysis.

Training was performed on a NVIDIA DGX-2 system with 16 Tesla V100 32-GB GPUs, dual Intel Xeon Platinum 8168 2.7-GHz 24- core CPUs, and 1.5 TB of system memory. Each training session was distributed across 4 GPUs using a mirrored strategy with synchronized global update variables. Up to four training sessions were run simultaneously using 16 total GPUs.

Qualitative Assessment: Multi-reader Study of Conventional vs. Synthetic Contrast Enhanced MRI

Thirty pairs of simulated MRI and real MRI from the test sets of two cross-validation folds were obtained for the multi-reader study. A breast radiology attending (A.L), with six years of experience and who was not a study reader, reviewed the real MRIs while blinded to the simulated MRIs. The radiologist identified 22 cases with an identifiable and measurable mass on the real MRI which were then included in the reader study. Cases without an identifiable index mass on the real MRI and cases with only foci and/or non-mass enhancement were excluded. Four blinded, breast fellowship-trained, MQSA-certified attending academic radiologists (authors B.N.J., K.M.R., T.K., J.H.H.) were selected as

readers. Reader experience in breast imaging, including interpretation of breast MRI, ranged from 3 years to 15 years. The 22 pairs of real and synthetic contrast MRI were separated, randomly assigned, and read between the two sessions, with each session including a mix of synthetic and conventional post-contrast images in randomized order. The two sessions were separated by a two-week memory washout period to reduce the influence of recall bias. All image metadata were removed and studies were assigned a randomly selected identification number. To ensure all readers evaluate and measure the same index lesion, the laterality and location of the index malignancy were provided to the readers. Readers were blinded to whether each study was a simulated MRI or real MRI. For each study the readers were asked: 1) "Does the image look like a real contrast-enhanced breast MRI?" (Yes/No), 2) "Is index tumor enhancement present?" (Yes/No). Readers provided a quantitative maximum dimension measurement of the dominant index mass at the site of the biopsy-proven malignancy. Readers also provided qualitative assessment of diagnostic image quality by scoring on a 5-point Likert-type scale: 5 = excellent (acceptable for diagnostic use), 4 = good (acceptable for diagnostic use), 3 = acceptable (acceptable for diagnostic use but with minor issues), 2 = poor (not acceptable for diagnostic use), or 1 = unacceptable (not acceptable for diagnostic use). For studies that were rated not acceptable for diagnostic use (score 1 or 2), readers indicated the reason: artifact, image resolution, lack of expected enhancement, or other (free text). Readers viewed and measured indexed tumors on an internet-based picture archiving and communication (https://www.pacsbin.com/). Reader responses were recorded in an Excel spreadsheet (version 16.64).

Quantitative Metrics for Image Comparison

Similarity between real and simulated MRIs were quantitatively evaluated using three similarity metrics four error metrics computed across the whole breast (15–17). Similarity metrics included normalized neighborhood cross correlation (CC) with a 5-voxel radius, histogram mutual information (MI) with 64 histogram bins, and the structural similarity index with a window size of 9 pixels. Error metrics were calculated after scaling normalized images to a common clinical MRI intensity range (12-bit or 0–4095) and included normalized root mean square error, symmetric mean absolute percent error, log accuracy ratio, and median symmetric accuracy. Similarity and error metrics were implemented in Python with Numpy 1.19.2, except for CC and MI, which were implemented with the Nipype 1.6.0 interface to Advanced Normalization Tools. Neighborhood-based similarity metrics were evaluated on images cropped to the tight bounding box of the whole breast. Metrics not dependent on voxel neighborhoods (all of the error metrics) were evaluated on flattened (one-dimensional) arrays of voxels within the evaluation region. Full code for network implementation is provided at https://github.com/ecalabr/breast_simulated_gad (specific version hash: 76e5347b5e9b0e98ab02db1b1728986412050de8).

Cases with Failed Tumor Enhancement

Of the 96 cases, seven demonstrated failed tumor enhancement. Breast density distribution among these seven cases was one scattered fibroglandular tissue, four heterogeneously dense, and one extremely dense. This distribution is similar to that in the entire patient sample (Table 1). These seven cases also included all levels of background parenchymal

enhancement (two minimal, three mild, one moderate, and one marked). Tumor size ranged from 0.8 to 3.3 cm and did not appear to be a major factor in failed enhancement prediction. Notably, one case of invasive lobular and one of invasive mucinous carcinoma demonstrated failed simulated tumor enhancement. This raises the possibility that cancer subtypes may exhibit different mathematical relationships between pre- and post-contrast series. The relatively small number of these less common tumor subtypes in our dataset may have limited algorithm modeling of their signal relationships. The five cases of invasive ductal carcinoma with failed tumor enhancement could be related to heterogeneity in tumor appearance and signal relationships within tumor subtypes as well. A larger dataset that includes robust numbers of less common tumor subtypes and various tumor appearances within each subtype may help improve algorithm generalizability and computational contrast simulation.

Relative Contribution of Input Series

The relative contribution of each of the pre-contrast input series to the simulated post-contrast images was determined by training the dCNN model on 15 MRIs with each combination of the T1-FS series plus one pre-contrast series (T1-FS + T1 nonFS, T1-FS + T2 FS, T1-FS + DWI, and T1-FS + apparent diffusion coefficient). Training for each combination consisted of predetermined 20 epochs without early stopping. Model performance of each combination was assessed using similarity and error metrics as described previously. 95% confidence intervals of similarity and error metrics were calculated. Statistical significance was determined by using Wilcoxon signed-rank test.

All pre-contrast models trained with combinations of two series (T1-FS + one pre-contrast series) demonstrated lower performance by similarity and error metrics than the full model; however, not all differences reached statistical significance (Supplemental Figures 2 and 3). In addition to the T1-FS images, T2-FS and b0 were the most important pre-contrast series for contrast simulation with T1-FS + T2-FS and T2-FS + b0 demonstrating performance metrics closest to the full model among the pre-contrast combinations.

Abbreviations:

FS fat-suppressed

References:

1. Comstock CE, Gatsonis C, Newstead GM, et al. Comparison of Abbreviated Breast MRI vs Digital Breast Tomosynthesis for Breast Cancer Detection among Women with Dense Breasts Undergoing Screening. JAMA - Journal of the American Medical Association. American Medical Association; 2020;323(8):746–756. doi: 10.1001/jama.2020.0572. [PubMed: 32096852]

2. Mann RM, Athanasiou A, Baltzer PAT, et al. Breast cancer screening in women with extremely dense breasts recommendations of the European Society of Breast Imaging (EUSOBI). European Radiology 2022. Springer; 2022;1–10. doi: 10.1007/S00330-022-08617-6.

3. Ramalho J, Semelka RC, Ramalho M, Nunes RH, AlObaidy M, Castillo M. Gadolinium-based contrast agent accumulation and toxicity: An update. American Journal of Neuroradiology. American Society of Neuroradiology; 2016. p. 1192–1198. doi: 10.3174/ajnr.A4615.

4. Yabuuchi H, Matsuo Y, Sunami S, et al. Detection of non-palpable breast cancer in asymptomatic women by using unenhanced diffusion-weighted and T2-weighted MR imaging: Comparison with

mammography and dynamic contrast-enhanced MR imaging. Eur Radiol. Springer; 2011;21(1):11–17. doi: 10.1007/s00330-010-1890-8. [PubMed: 20640898]

5. McDonald ES, Hammersley JA, Chou SHS, et al. Performance of DWI as a rapid unenhanced technique for detecting mammographically occult breast cancer in elevated-risk women with dense breasts. American Journal of Roentgenology. American Roentgen Ray Society; 2016;207(1):205–216. doi: 10.2214/AJR.15.15873. [PubMed: 27077731]

6. Kazama T, Kuroki Y, Kikuchi M, et al. Diffusion-weighted MRI as an adjunct to mammography in women under 50 years of age: An initial study. Journal of Magnetic Resonance Imaging. John Wiley & Sons, Ltd; 2012;36(1):139–144. doi: 10.1002/jmri.23626. [PubMed: 22359367]

7. Trimboli RM, Verardi N, Cartia F, Carbonaro LA, Sardanelli F. Breast Cancer Detection Using Double Reading of Unenhanced MRI Including T1-Weighted, T2-Weighted STIR, and Diffusion-Weighted Imaging: A Proof of Concept Study. American Journal of Roentgenology. American Roentgen Ray Society; 2014;203(3):674–681. doi: 10.2214/AJR.13.11816.

8. Telegrafo M, Rella L, Stabile Ianora AA, Angelelli G, Moschetta M. Unenhanced breast MRI (STIR, T2-weighted TSE, DWIBS): An accurate and alternative strategy for detecting and differentiating breast lesions. Magn Reson Imaging. Elsevier Inc.; 2015;33(8):951–955. doi: 10.1016/j.mri.2015.06.002. [PubMed: 26117691]

9. Kang JW, Shin HJ, Shin KC, et al. Unenhanced magnetic resonance screening using fused diffusion-weighted imaging and maximum-intensity projection in patients with a personal history of breast cancer: role of fused DWI for postoperative screening. Breast Cancer Res Treat. Springer New York LLC; 2017;165(1):119–128. doi: 10.1007/s10549-017-4322-5. [PubMed: 28577079]

10. Guo Y, Cai YQ, Cai ZL, et al. Differentiation of clinically benign and malignant breast lesions using diffusion-weighted imaging. J Magn Reson Imaging. J Magn Reson Imaging; 2002;16(2):172–178. doi: 10.1002/JMRI.10140. [PubMed: 12203765]

11. Narayana PA, Coronado I, Sujit SJ, Wolinsky JS, Lublin FD, Gabr RE. Deep Learning for Predicting Enhancing Lesions in Multiple Sclerosis from Noncontrast MRI. Radiology. Radiological Society of North America Inc.; 2020;294(2):398–404. doi: 10.1148/radiol.2019191061.

12. Kleesiek J, Morshuis JN, Isensee F, et al. Can Virtual Contrast Enhancement in Brain MRI Replace Gadolinium? Invest Radiol. Lippincott Williams and Wilkins; 2019;54(10):653–660. doi: 10.1097/RLI.0000000000000583. [PubMed: 31261293]

13. Sun H, Liu X, Feng X, et al. Substituting Gadolinium in Brain MRI Using DeepContrast. Proceedings - International Symposium on Biomedical Imaging. IEEE Computer Society; 2020;2020-April:908–912. http://arxiv.org/abs/2001.05551. Accessed January 2, 2021.

14. Calabrese E, Rudie JD, Rauschecker AM, Villanueva-Meyer JE, Cha S. Feasibility of Simulated Postcontrast MRI of Glioblastomas and Lower-Grade Gliomas by Using Three-dimensional Fully Convolutional Neural Networks. https://doi.org/101148/ryai2021200276. Radiological Society of North America; 2021;3(5):e200276. doi: 10.1148/RYAI.2021200276.

15. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage. Neuroimage; 2011;54(3):2033–2044. doi: 10.1016/j.neuroimage.2010.09.025. [PubMed: 20851191]

16. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing. 2004;13(4):600–612. doi: 10.1109/TIP.2003.819861. [PubMed: 15376593]

17. Morley SK, Brito T v., Welling DT. Measures of Model Performance Based On the Log Accuracy Ratio. Space Weather. Blackwell Publishing Ltd; 2018;16(1):69–88. doi: 10.1002/2017SW001669.

18. Myles PS, Cui JI. Using the Bland–Altman method to measure agreement with repeated measures. Br J Anaesth. Elsevier; 2007;99(3):309–311. doi: 10.1093/BJA/AEM214. [PubMed: 17702826]

19. R: The R Project for Statistical Computing https://www.r-project.org/. Accessed December 14, 2021.

20. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods 2007 39:2. Springer; 2007;39(2):175–191. doi: 10.3758/BF03193146.

21. JK O BEM, SD H, JS O. Breast cancer tumor size: correlation between magnetic resonance imaging and pathology measurements. Am J Surg. Am J Surg; 2008;196(6):844–850. doi: 10.1016/J.AMJSURG.2008.07.028. [PubMed: 19095098]

22. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage. Neuroimage; 2011;54(3):2033–2044. doi: 10.1016/J.NEUROIMAGE.2010.09.025. [PubMed: 20851191]

23. Calabrese E, Villanueva-Meyer JE, Cha S. A fully automated artificial intelligence method for non-invasive, imaging-based identification of genetic alterations in glioblastomas. Scientific Reports 2020 10:1. Nature Publishing Group; 2020;10(1):1–11. doi: 10.1038/s41598-020-68857-8.

24. ITK-SNAP Home. http://www.itksnap.org/pmwiki/pmwiki.php. Accessed December 26, 2020.

25. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag; 2015. p. 234–241. doi: 10.1007/978-3-319-24574-4_28.

26. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. http://image-net.org/challenges/LSVRC/2015/. Accessed January 2, 2021.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Key Results:

- In a retrospective study of 96 women with invasive breast cancer, simulated contrast-enhanced breast MRIs were quantitatively similar to real contrast-enhanced MRIs with a mean structural similarity index 0.88 ± 0.05.

- Simulated and real contrast-enhanced MRIs demonstrated high degree of enhancing tumor overlap (Dice coefficient 0.75 ± 0.25).

- Breast radiologists assessed all simulated MRIs as having the appearance of real contrast-enhanced MRIs and almost all were of diagnostic quality (84 of 88; 95%).
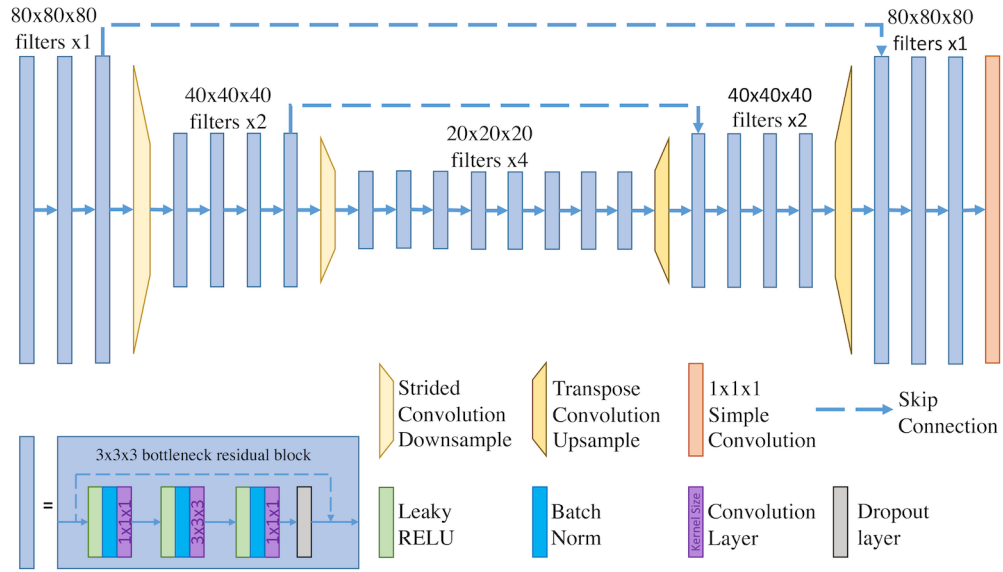
Figure 1:

Schematic of the Deep Learning, Fully Convolutional Neural Network Architecture

The deep learning, convolutional neural network was trained on 80×80×80 voxels training patches from preprocessed images. Our network architecture consisted of a deep learning, convolutional neural network with three-dimensional convolution bottleneck residual blocks (blue blocks), strided convolution downsampling (yellow trapezoids), transpose convolution upsampling (orange trapezoids), and long-range skip connections with feature concatenation (dashed lines). A 1×1×1 convolutional layer (orange block) mapped features to final output image patches. A schematic of the 3×3×3 bottleneck residual block is included (bottom left).
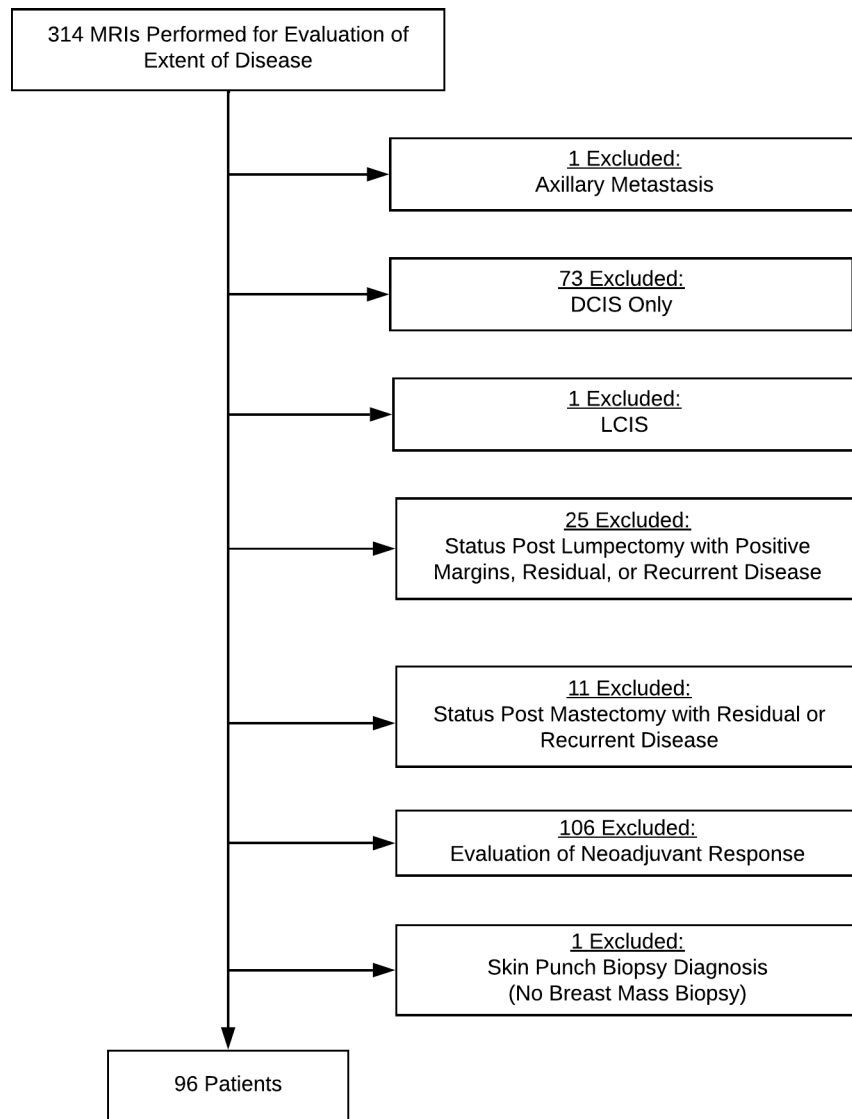
Figure 2:
Study Flow Chart
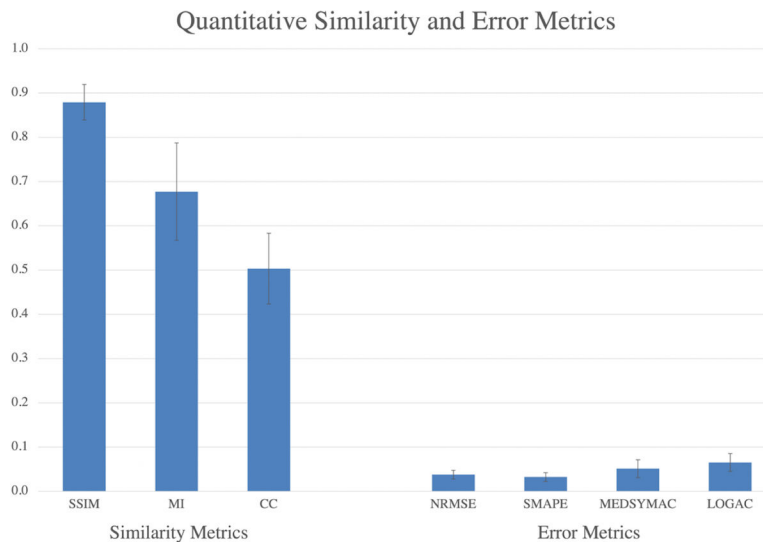Ductal carcinoma in situ (DCIS), lobular carcinoma in situ (LCIS).

Figure 3:

Quantitative Similarity and Error Metrics for Real versus Simulated Contrast-enhanced MRI across the Whole Breast

Overall, there was strong similarity and low voxel-wise error across the whole breast. Similarity metrics: structural similarity index (SSIM), histogram mutual information (MI), and normalized neighborhood cross correlation (CC). Four error metrics: normalized root mean square error (NRMSE), symmetric mean absolute percent error (SMAPE), median symmetric accuracy (MdSA), and log accuracy ratio (LOGAC).
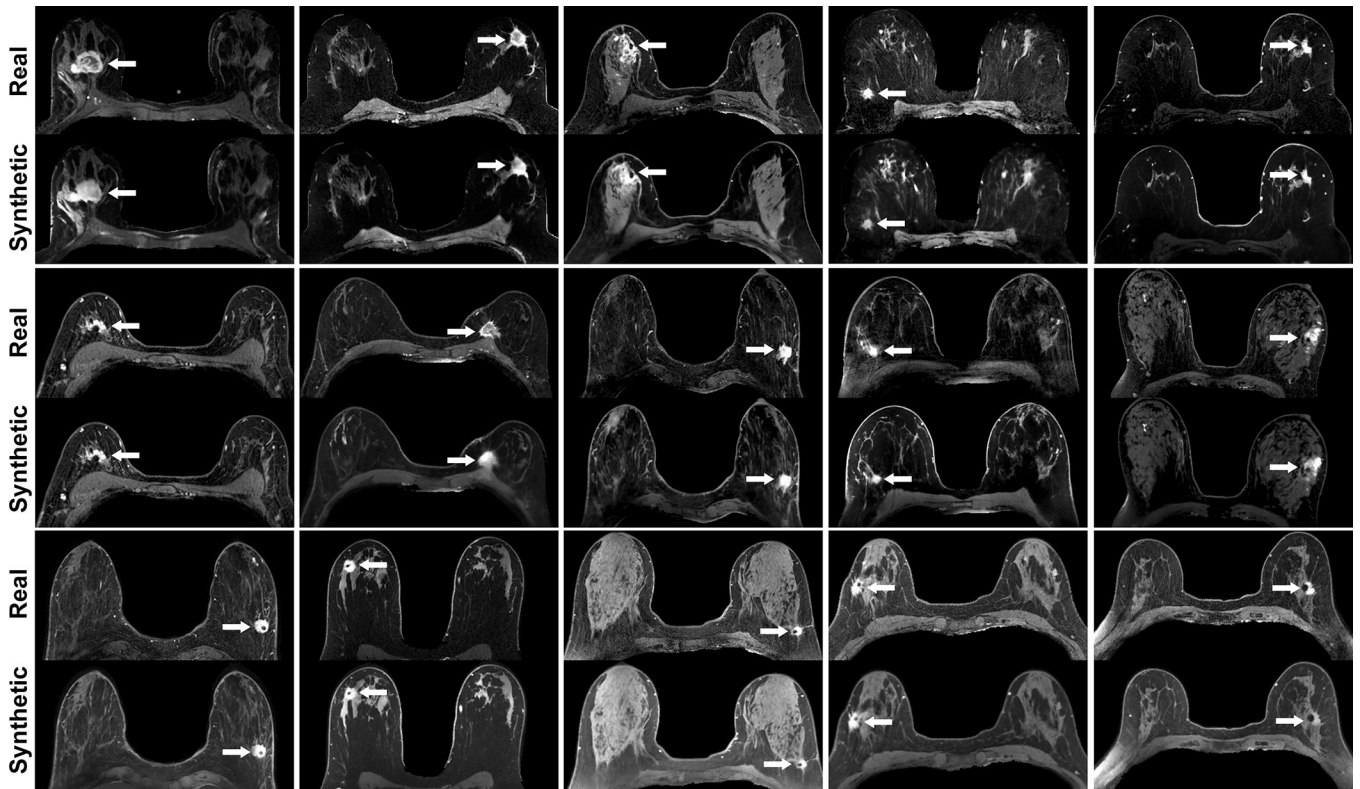
Figure 4.
Real versus Simulated Contrast-enhanced T1-weighted Axial Breast MRIs of Patients with Invasive Breast Cancer

Pairs of real (top) and simulated (bottom) contrast-enhanced breast MRI from 15 patients with invasive breast cancer (arrows). Intrathoracic and extramammary structures were masked in all images.

Figure 5:
Failed Enhancement of the Index lesion on the Simulated Contrast-enhanced MRI
Simulated contrast-enhanced MRI (bottom) demonstrated failed index lesion enhancement
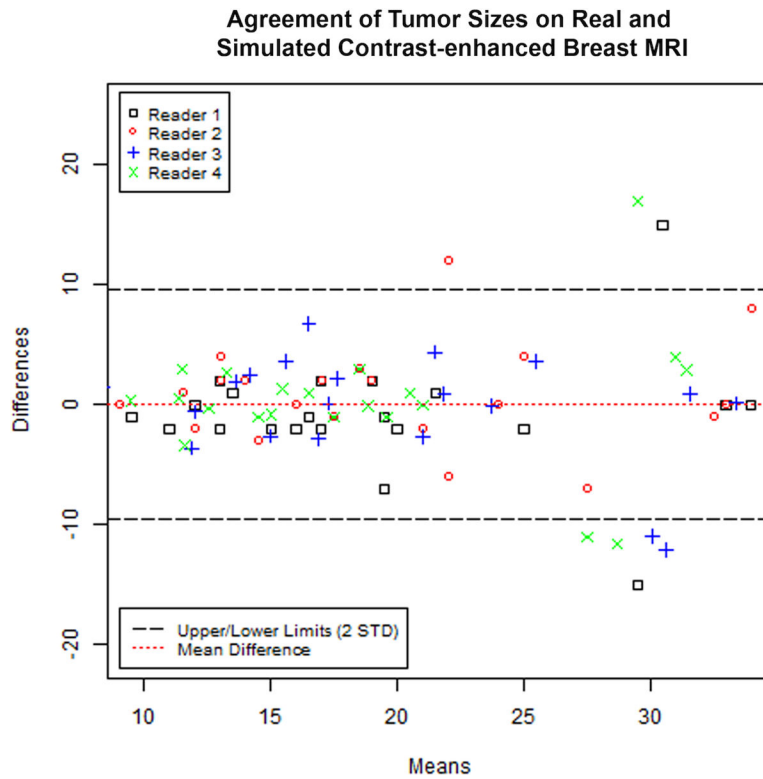(arrows) compared with the real contrast-enhanced MRI (top).

Figure 6.
Agreement of Tumor Sizes on Real and Simulated Contrast-enhanced Breast MRI
Modified Bland-Altman plot of agreement of tumor sizes on real versus simulated contrast-enhanced breast MRI. The dotted red lines and black dashed lines represent bias and limits of agreement lines (two standard deviations above and below the mean), respectively.

Table 1:

Patient Clinical Characteristics

Characteristic	Whole Sample
No. of women	96
Age (y)	
Mean	52 ± 12
Range	26–76
Menopausal Status	
Premenopausal	47 (49%)
Postmenopausal	49 (51%)
Fibroglandular Tissue	
Fatty	0 (0%)
Scattered	27 (28%)
Heterogenous	43 (45%)
Extremely Dense	26 (27%)
Background Parenchymal Enhancement	
Minimal	15 (16%)
Mild	45 (47%)
Moderate	21 (22%)
Marked	15 (16%)
MRI Tesla	
1.5 T	19
3 T	77

Note: Data are numbers of women with percentages in parentheses, unless otherwise stated. Mean age is ± standard deviation.

*
Data in parentheses are interquartile range.

Table 2.

Tumor Characteristics

Characteristic	Whole Sample
ER Status	
Positive	19 (20%)
Negative	76 (79%)
Not Available	1 (1%)
PR Status	
Positive	38 (40%)
Negative	56 (58%)
Not Available	2 (2%)
HER2 Status	
Positive	75 (78%)
Negative	19 (20%)
Not Available	2 (2%)
Ki-67 Status	
<14%	46 (48%)
14%	42 (44%)
Not Available	8 (8%)
Nuclear Grade	
1	5 (5%)
2	46 (48%)
3	44 (46%)
Not Available	1 (1%)
Kinetics: Initial Phase	
Slow	0 (0%)
Medium	4 (4%)
Fast	92 (96%)
Kinetics: Delayed Phase	
Persistent	7 (7%)
Plateau	18 (19%)
Washout	71 (74%)
Type of Lesion	
Mass Only	51 (53%)
NME Only	8 (8%)
Mass and NME	37 (39%)
Lesion Size	
Mean	2.4 cm

Characteristic	Whole Sample
Median	2.0 cm
Range	0.7 cm – 7.8 cm

Table 3:

Multi-reader Assessment of Image Quality of Real versus Simulated Contrast-enhanced MRI

Image Quality	Real Post Contrast	Simulated Contrast
Combined (n=88)		
Non-diagnostic (unacceptable/poor)	0 (0%)	4 (5%)
Diagnostic (good/acceptable/excellent)	88 (100%)	84 (95%)
Combined (n=88)		
Unacceptable	0 (0%)	0 (0%)
Poor	0 (0%)	4 (5%)
Acceptable	12 (14%)	30 (34%)
Good	29 (33%)	38 (43%)
Excellent	47 (53%)	16 (18%)
Reader 1 (n=22)		
Unacceptable	0 (0%)	0 (0%)
Poor	0 (0%)	1 (5%)
Acceptable	1 (5%)	3 (14%)
Good	4 (18%)	9 (41%)
Excellent	17 (77%)	9 (41%)
Reader 2 (n=22)		
Unacceptable	0 (0%)	0 (0%)
Poor	0 (0%)	0 (0%)
Acceptable	6 (28%)	13 (59%)
Good	8 (36%)	7 (32%)
Excellent	8 (36%)	2 (9%)
Reader 3 (n=22)		
Unacceptable	0 (0%)	0 (0%)
Poor	0 (0%)	3 (14%)
Acceptable	3 (14%)	6 (27%)
Good	8 (36%)	9 (41%)
Excellent	11 (50%)	4 (18%)
Reader 4 (n=22)		
Unacceptable	0 (0%)	0 (0%)
Poor	0 (0%)	0 (0%)
Acceptable	2 (9%)	8 (36%)
Good	9 (41%)	13 (59%)
Excellent	11 (50%)	1 (5%)