T-SYNTH: A Knowledge-Based Dataset of Synthetic Breast Images

Christopher **Wiedeman**∗, Anastasiia **Sarmakeeva**∗, Elena **Sizikova,** Daniil **Filienko,** Miguel **Lago,** Jana G. **Delfino,** Aldo **Bdadano,**

Office of Science and Engineering Laboratories
Center for Devices and Radiological Health
U.S. Food and Drug Administration
Silver Spring, MD 20993 USA
∗ - These authors contributed equally to this work.

Abstract

One of the key impediments for developing and assessing robust medical imaging algorithms is limited access to large-scale datasets with suitable annotations. Synthetic data generated with plausible physical and biological constraints may address some of these data limitations. We propose the use of physics simulations to generate synthetic images with pixel-level segmentation annotations, which are notoriously difficult to obtain. Specifically, we apply this approach to breast imaging analysis and release T-SYNTH, a large-scale open-source dataset of paired 2D digital mammography (DM) and 3D digital breast tomosynthesis (DBT) images. Our initial experimental results indicate that T-SYNTH images show promise for augmenting limited real patient datasets for detection tasks in DM and DBT. Our data and code are publicly available at: `https://github.com/DIDSR/tsynth-release`.

Keywords

Digital Breast Tomosynthesis (DBT); Synthetic Data; Lesion Detection

1. Background

Responsible for approximately two million new cases and over six hundred thousand deaths in 2022 alone (Sung et al., 2021), breast cancer remains a prominent global health concern, and is expected to account nearly one-third of all newly diagnosed cancers among women in the United States (DeSantis et al., 2016). According to the most recent report from International Agency for Research on Cancer (Bray et al., 2024), it is one of the most widespread cancers diagnosed worldwide, both in the number of cases and associated deaths. Consequently, medical imaging techniques are indispensable for screening, diagnosis, and further research into the disease. Historically, the most common imaging technique for breast cancer screening is digital mammography (DM), in which a 2D x-ray projection of a compressed breast is taken. Digital breast tomosynthesis (DBT), a pseudo-3D imaging technique, has been increasingly adopted, demonstrating improved screening performance (Asbeutah et al., 2019; Sprague et al., 2023).

The development of algorithms for segmenting, detecting, or classifying structures of interest (e.g., cancerous lesions) within these images is desirable for improving both patient outcomes as well as workflow efficiency. A limiting factor to developing data-driven algorithms, however, is the access to representative datasets with appropriate annotations. Collecting representative medical imaging data is challenging due to cost, patient privacy regulations, and other constraints, but obtaining appropriate annotations is even more taxing, since annotators must have specialized domain knowledge. The labeling process itself is laborious due to the high resolution and fine details typically present in these images. Very few DM and DBT datasets are publicly available. Furthermore, the cohort characteristics and amount of information included in these datasets are highly variable, which limits the robustness of their application. Only a fraction of publicly-available datasets contain
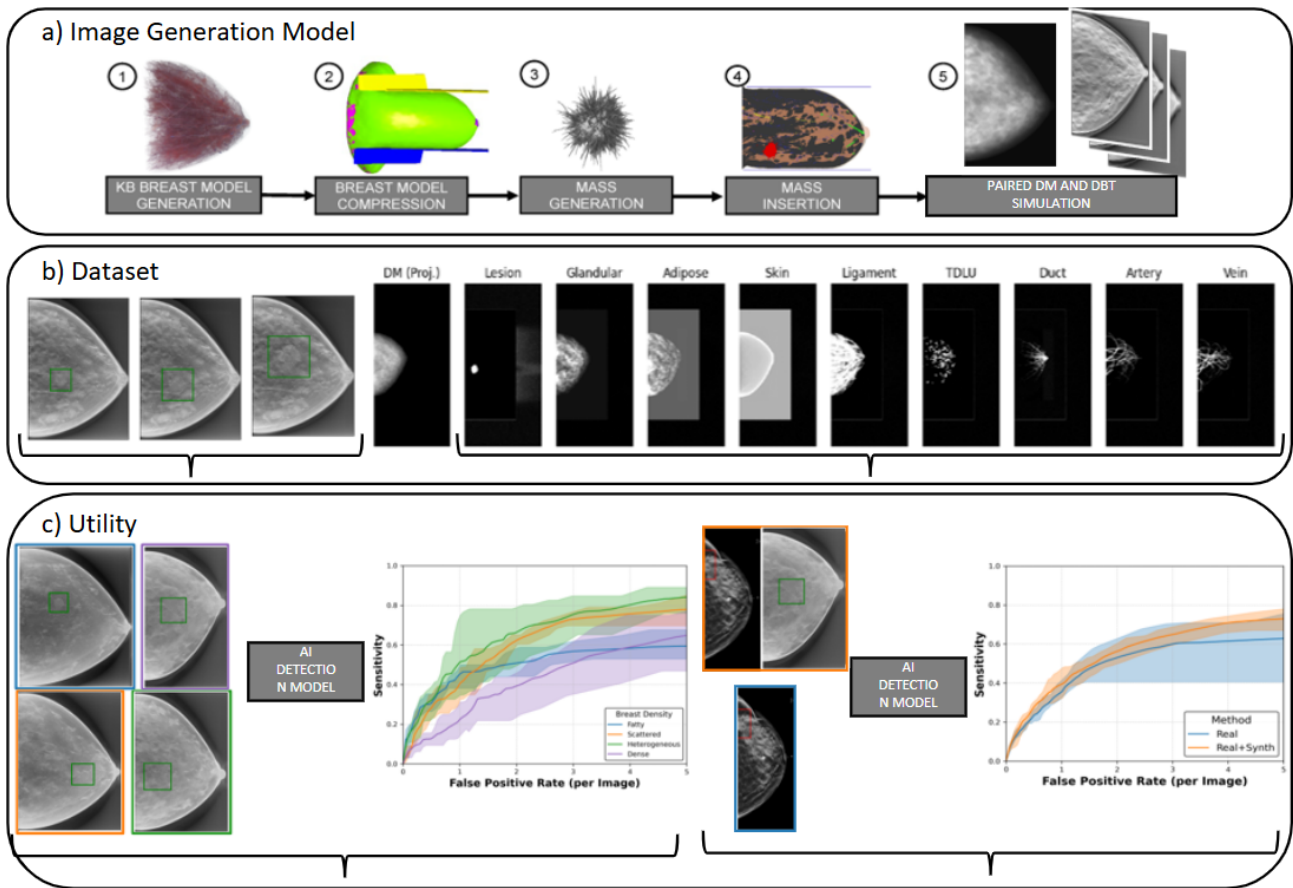
Figure 1: Summary of T-SYNTH: a) Knowledge-based process for simulating paired synthetic digital mammography (DM) and digital breast tomosynthesis (DBT) examples. b) Synthetic T-SYNTH data, which includes varying lesion sizes, lesion densities (left), and tissue masks (right). c) Utility of T-SYNTH, which includes comparing performance of AI models for lesion detection in different subgroups (left) and augmenting existing training data with T-SYNTH for better detection in patient breast images (right).

pixel-level annotations for relevant tissues, since pixel-level annotations typically require a time-consuming, costly, and laborious annotation process conducted by a specialist.

Breast Image Analysis Tasks Most AI applications in breast image analysis focus on the task of mass detection (annotating bounding boxes over regions of interest) and segmentation. Mass detection is often studied because of its direct applicability in Computer Aided Diagnosis (CAD) style software, because of the improved diagnostic capabilities provided by such models. Mass detection and segmentation in DM and DBT images is especially valuable for reducing clinical workloads (Zhang et al., 2023; Raya-Povedano et al., 2021). Mass detection and segmentation tasks often occur together, since segmentation models may contain a separate bounding box detection step, where any predicted regions of interest (ROIs) are classified by lesion presence/properties and then further segmented by the model, as in a Mask R-CNN (He et al., 2017). Alternatively, a model such as a U-Net can segment findings directly (Ronneberger et al., 2015). Regardless of the model used, AI techniques have been applied to the task of segmentation

and detection in both DM Salama and Aly (2021) and DBT modalities (Bai et al., 2021), albeit with varying success, occasionally being prone to false-positive predictions or over-fitting (Zhang et al., 2023). AI analysis of non-mass breast tissues, such as fibroglandular and adipose tissues, remains under-explored, despite the utility of these tissues to identify potentially cancerous regions (Zhu et al., 2017; Tiryaki and Kaplanoğlu, 2022), leading to more precise diagnostics (Lam et al., 2000). To the best of our knowledge, there exist only a few works in segmenting non-lesion breast tissues (Yamamuro et al., 2022), and among them, only two use deep learning methods (Tiryaki and Kaplanoğlu, 2022; Menegatti Pavan et al., 2016). This scarcity is due, at least in part, to the severe dearth of annotations for these tissues in publicly available datasets (see Appendix for a summary of existing datasets).

Synthetic Data A potential way to mitigate data scarcity and class imbalance is to utilize synthetic data generation. Employing generative AI to produce these data, however, poses several shortcomings. First, the synthesized ground truth information is limited by what is captured in the ob-

served data (Badano et al., 2023). Limitations stem from both the imaging system (for example, tissue orientations that cannot be differentiated because they lie along the system's null space) and human annotators (imperfect segmentation masks or bounding boxes). Secondly, generative models can perpetuate biases that exist within the data, marginalizing underrepresented demographics and fringe cases (Zack et al., 2024). Ironically, synthetic data are most needed in instances where these biases exist. Finally, even when data restrictions are ignored, the finite expressivity and imperfect training process of the model itself results in considerable errors. For example, Kelkar et al. (2023) found that generative adversarial networks trained on simulated medical images failed to match their training sets in several key clinically relevant features. A recent work by Shumailov et al. (2024) also demonstrated that generative models trained over generations of synthetic data compounded errors and led to 'model collapse'.

Our dataset relies on knowledge-based (KB) models which inherently incorporate domain expertise ("knowledge" of physics and biology), which are used to procedurally create realistic images and annotations. Although KB models require rigorous domain-specific knowledge and intensive, realistic simulation, the pre-image ground-truth characteristics of a sample (for example, the underlying breast phantom) are precisely known. This allows for subgroup analyses, controlled comparison of different scan procedures, and control of dataset balance. Here, for example, we demonstrate how our KB data can analyze detection performance among different mass sizes and densities, information that is not available from patient data.

2. Summary

In this project, we demonstrate how knowledge-based (KB) models can be used as a source of synthetic data containing bounding box and pixel-level annotations for breast tissues in DM and DBT. We publicly release this dataset to accelerate AI development in medical imaging applications. The model we use, developed by Graff (2016), consists of multiple components (glandular and adipose tissue, skin, ligaments, ducts, veins and arteries), and allows the insertion of lesions (see Fig. 1). By manipulating the properties of each component during the virtual rendering process, we can generate a variety of richly annotated examples. This ability is unique to the data generation models that decouple multi-scale components of the patient anatomy and sensor, unlike generative approaches that learn from imaging data alone Badano et al. (2023) (see Sec. 1 for more discussion). Our contributions are:

- We release T-SYNTH, a public synthetic dataset of paired DM (2D imaging) and DBT (3D imaging) images derived from a KB model, with pixel-level segmentation

and bounding boxes of a variety of breast tissues.

- We demonstrate how T-SYNTH can be used for subgroup analysis. Specifically, Faster-RCNN Ren et al. (2015) is trained for and evaluated for lesion detection in a balanced dataset; results reveal expected trends in subgroup performance in both DM and (C-View) DBT (e.g., less dense lesions are harder to detect).

- We train detection models on limited patient data in both DM and DBT (C-View), and show that augmenting training data with T-SYNTH can improve performance.

3. Methods

The open-source Virtual Imaging Clinical Trials for Regulatory Evaluation (VICTRE) pipeline (Badano et al., 2018) is used to simulate DM and DBT images with variations to breast density and mass properties, similar to the M-SYNTH dataset (Sizikova et al., 2022). The resulting datasets can be used for downstream AI analysis as discussed in Sec. 4.1.

3.1 Phantom Generation

Controlled settings and their factors for phantom generation include: breast density: fatty, scattered, heterogeneous, and dense; lesion density: 1.0, 1.06, and 1.1 times the density of glandular tissue and lesion diameter: 5, 7, and 9 mm. 150 breast phantoms with lesions were generated for each factorial combination except for 9 mm lesions in heterogeneous and dense breasts, due to the smaller size of these denser breasts. Corresponding negative (lesion absent) samples were also generated for each positive sample, for a total of 9,000 images (4,500 lesion present and 4,500 lesion absent).

Lesion Growth and Insertion Sengupta et al. (2024) introduced a computational model for lesion growth of breast cancer within a three-dimensional voxelized breast model, accounting for stiffness of surrounding structures and considering both avascular and vascular phases of tumor development. The growth of lesions simulated with VICTRE pipeline. During the avascular phase, the tumor growth is limited by nutrient diffusion from surrounding tissues, leading to high interstitial pressure at the core and eventual necrosis due to oxygen deprivation. This triggers the release of tumor angiogenesis factors, promoting new blood vessel formation and transitioning the tumor into the vascular phase.

The growth dynamics are modeled using advection-reaction-diffusion equations. The model calculates interstitial tumor pressure and incorporates oxygen as a primary nutrient, influencing cell metabolism, angiogenesis, and metastasis. The tumor cells' proliferation is guided by local tissue stiffness, with softer tissues offering less resistance,

thereby promoting tumor expansion in these regions. For imaging, the generated lesion models are incorporated into fatty breast models using voxel replacement techniques. The approach to modeling breast cancer lesion growth in silico, enabling the simulation of realistic and biologically-relevant lesion morphologies. Due to the model's ability to account for the mechanical properties of surrounding tissues, we used it in our research.

3.2 Acquisition System Modeling

DBT Imaging DBT images are created via a replica of the Siemens detector based on MC-GPU (Badal and Badano, 2009). DM images are produced using the GPU-accelerated Monte Carlo x-ray transport simulation code, replicating the Siemens Mammomat Inspiration system (matching those from the M-SYNTH work).

C-View Images Majority of existing patient datasets are limited to C-View images (i.e., 2D images obtained from DBT projections) as they are less computationally intensive to analyze. We therefore perform comparative evaluation of T-SYNTH using lesion detection in C-View. Commercial algorithms for synthesizing C-View images from DBT data are proprietary. As such, we used the C-View approximation from Klein et al. (2023), which applies a spherical sharpening filter to the 3D volume and then taking the pointwise maximum along the compressed axis. For all instances, lesion bounding boxes were computed to encapsulate positive pixels in the known lesion mask.

4. Experimental Setup

We investigate T-SYNTH as tool for subgroup analysis in synthetic data (Sec. 5.2), augmenting limited patient datasets and investigations into how the composition of data affect downstream performance (Sec. 5.3). The Torchvision implementation of Faster R-CNN was used as the detection model, with a ResNet50 feature pyramid network pretrained on COCO (Ren et al., 2015). Five model instances each for DBT C-View and DM were trained and evaluated. Training loss was the summation of the classifier/objectness and box regression losses from the region proposal network and R-CNN. AdamW optimizer with a learning rate of 1e-4 was used, with a batch size of 24 images. Models were tuned by measuring sensitivity on the validation set at a decision threshold of 0.5, which was checked every 100 training steps. Single V100s were used to train models through 3000 training steps for experiments with subsets of data and on 5000 training steps when we trained on all data, which took about 5-8 hours to complete. We follow the conventions set in M-SYNTH and analyze models trained on synthetic data and patient data, as outlined below. Experiments are conducted for both DM and C-View DBT.

Breast Density	Train Set	Val Set	Test Set	Total
Fatty	900/900	225/225	225/225	1350/1350
Scattered	900/900	225/225	225/225	1350/1350
Hetero	600/600	150/150	150/150	900/900
Dense	600/600	150/150	150/150	900/900
Total	**3000/3000**	**750/750**	**750/750**	**4500/4500**

Table 1: Distribution of **image** samples $(+/-)$ from the T-SYNTH dataset (DM and DBT).

4.1 Datasets

Synthetic Subgroup Analysis The subgroup factors we analyze for detection performance are breast density (4 levels), lesion density (3 levels) and lesion size (3 levels), using 300 distinct digital models (phantoms) for each possible subgroup, see Tab. 1. Note that unlike patient data which is unbalanced in positive and negative samples, T-SYNTH is balanced.

Patient Data Augmentation To explore data augmentation using T-SYNTH, five detection models were trained using only patient data, and five were trained with patient + synthetic data. EMBED was used as the patient (real) dataset Jeong et al. (2023). Unfortunately, very few DBT (all C-View) images of fatty and dense breasts with findings are present in EMBED data. The available images were split into a training, validation and test set, such that each breast density was approximately represented equally in the validation and test sets, and no patients were shared between sets (see Appendix). Although a greater number of DM images are available, we maintained the same split for DM experiments for consistency.

Aside from the data used, all other training parameters are identical to the synthetic only training. A batch size of 26 images was used; 10 of which were from T-SYNTH in the patient + synth trials. Both methods had access to the entire patient training set. Patient images were sampled such that half contained at least one finding in each batch.

4.2 Evaluation

Each model is evaluated by calculating the number of true positives (successfully detected lesions), false positives (regions falsely declared as a lesion), and false negatives (missed lesions) that a model outputs over the test set at varying decision thresholds. Similar to the DBTex detection challenge for DBT (Konz et al., 2023), a true positive was scored if the distance between the prediction and a ground truth boxes' centers was less than either half of the diagonal length of the ground truth box or 100 pixels. Also similar to this challenge, model sensitivity is reported in relation to the average number of false positives per image as a free-response receiver operator characteristic (FROC) curve. FROC is often preferred to precision-recall in medical screening contexts, where most exams do not contain any

true findings, and the number of false positives per image more directly indicates the clinical workload incurred by the system. Finally, for each experiment, we retrain each model five times with different random seeds, and report the mean, minimum and maximum values of sensitivity at each FPR rate across the five models, respectively. Finally, differences in performance are analyzed across breast density, as well as lesion size and density subgroups.

5. Results and Discussions

5.1 T-SYNTH Data Visualization

Sample T-SYNTH images are shown in Fig. 2. Visual differences can be observed between synthetic DM, DBT, and C-View images. These differences largely stem from the distinct reconstruction algorithms used, which notably differ from those employed with patient clinical data. Despite these variations, consistency in the underlying patient anatomy remains evident across image types.

Clinically relevant and expected trends related to lesion visibility can also be seen. Specifically, lesions are less distinct from the background fibroglandular tissue in higher breast density categories. This finding aligns with established medical literature, which notes that dense tissue regions can either obscure lesions, particularly projections with overlapping tissues, or mimic them. In the highest breast density considered (last row of Fig. 2), DBT imaging provides a clearer visualization of lesions that were otherwise obscured by dense tissue in traditional DM images.

Similar visibility trends were observed across additional subgroup characteristics (see Appendix): smaller lesions consistently proved more difficult to detect, whereas smaller but denser lesions were relatively easier to identify. Consequently, T-SYNTH offers a robust paired synthetic DM and DBT dataset whose visual trends align closely with expectations derived from existing medical research. This consistency enhances our confidence in the dataset's validity and its potential applicability for training and evaluation purposes in breast imaging research.

In Fig. 3, we qualitatively compare lesions in the EMBED and T-SYNTH datasets. We find that the EMBED bounding boxes may contain multiple lesions and may not be as consistently labeled as the bounding boxes in the T-SYNTH dataset.

5.2 T-SYNTH Subgroup Analysis

In Fig. 4, we report detection results using FROC curves on models trained on synthetic data and evaluated on different subgroups in T-SYNTH. The x-axis represents the False Positive Rate (per image), and the Y-axis shows Sensitivity (True Positive Rate) defined as the proportion of correctly detected ground truth lesions. To generate
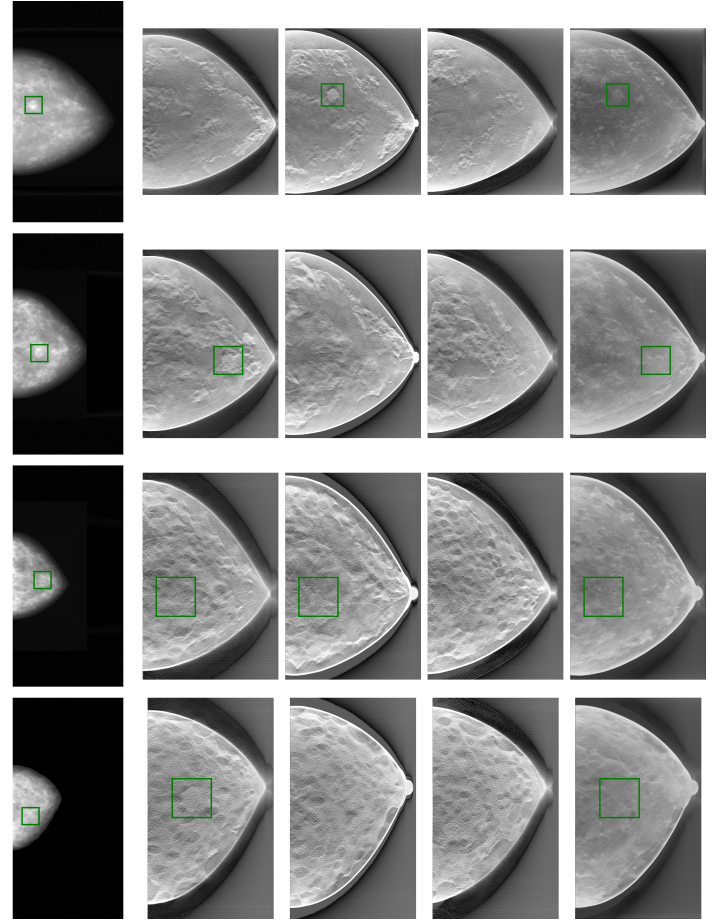

Figure 2: Example T-SYNTH images with varying breast densities (row 1: fatty, row 2: scattered, row 3: heterogeneous, row 4: dense). Column 1: DM, columns 2-4: DBT slices at quarter, half, and three-quarters of the total volume depth, Column 5: DBT C-View.

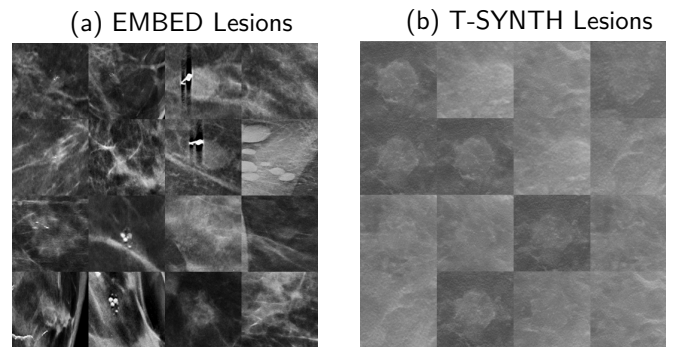(a) EMBED Lesions   (b) T-SYNTH Lesions

Figure 3: Feature analysis of lesions in the EMBED and T-SYNTH datasets.

these curves, each prediction was processed across a range of score thresholds (determined via the validation set) to compute detection outcomes. A detection was counted as a true positive if it matched a ground truth lesion and had not been matched previously by a higher-scoring prediction. Detections that did not match any ground truth lesions were counted as false positives. For each threshold, the total
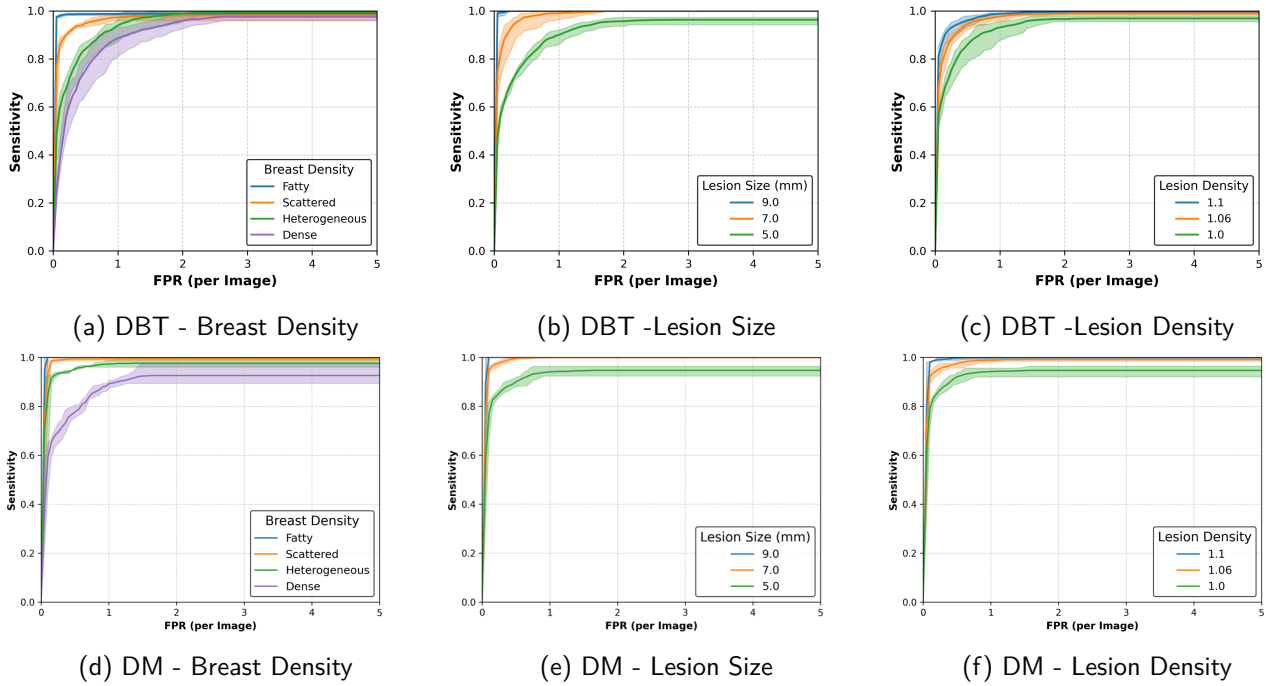
Figure 4: Synthetic Only Detection Results (Free response ROC Curve) comparing performance across subgroups: breast density, lesion size and density.
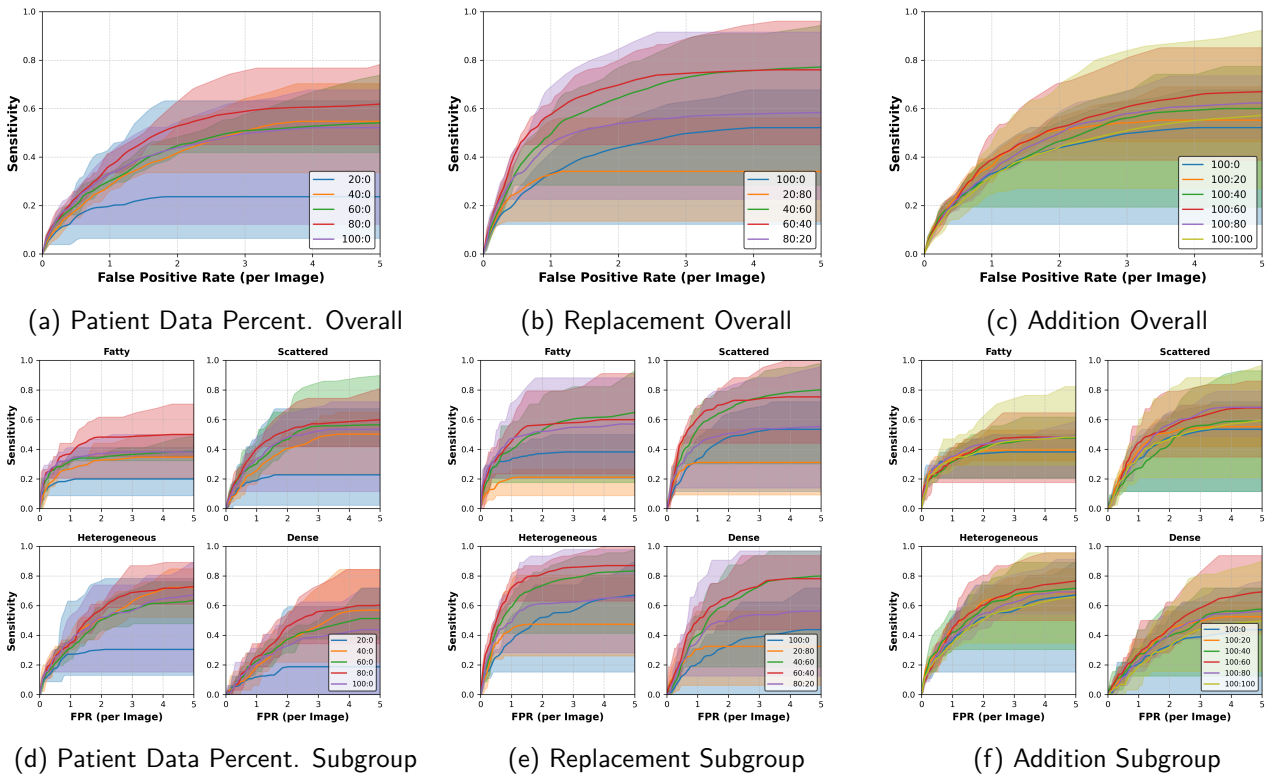

Figure 5: *Patient data percentage*: evaluation of when only a fraction of patient training data is available. *Replacement*: evaluation of patient data with replacement of synthetic examples. *Addition*: evaluation of patient data and addition synthetic examples

number of true positives was divided by the total number of ground truth lesions to obtain sensitivity (Se), while false positives (FP) were averaged per image. The shaded regions indicate the range (minimum to maximum) of sensitivity values across readers. In both DBT and DM examples, we find that performance improves with decreased density of fibroglandular tissue (i.e., breast density), increased lesion size, and increased lesion density, confirming the visual trends (see Sec. 5.1) and the overall clinical expectation. The spread of minimum values in DBT C-View images was slightly higher than in DM, likely due to additional noise. Because the analyzed subgroups are balanced and the nature of the phantoms control for other confounding factors, we can isolate these variables as the cause of differences in performance, boasting a potential advantage of T-SYNTH over patient data for subgroup analysis.

5.3 Impact of Data Composition

In this section, we evaluate the effect of data composition (patient:synthetic) examples on performance. In contrast to the above, we select a random subset of patient data to ensure that the number of patient and synthetic examples for training is fixed, but the composition of the training data changes. In both patient and synthetic cases, 100% of data indicates the inclusion of 800 examples: 400 with lesions and 400 with no lesions. The test set here consists of 800 examples. As before, the FROC curves include five models, where solid line is average among readers and shaded area is shows the minimum and maximum, respectively. We run three types of experiments:

- **Patient data percentage**: when model has only a fraction (20%, 40%, .., 100%) of patient training data available (Fig. 5a and Fig. 5d).

- **Replacement**: when a proportion of patient data is replaced with synthetic data, e.g., patient (20%) + synth (80%) (Fig. 5b and Fig. 5e).

- **Addition**: when a proportion of synthetic data is added to the full available patient training set, e.g., patient (100%) + synth (20%) (Fig. 5c and Fig. 5f).

From the patient data percentage experiment, we find that average performance drops significantly when only 20% of patient data is available, but other experiments with different data percentage 40%, 60% and 80% perform in a similar range. For fatty and scattered breasts, there is a drop from using only 40% of data, but for dense and heterogeneous breasts, there is not. When a proportion of patient data is replaced with synthetic data and compared with the case when all patient data is present (100:0), we find that replacing 40% to 80% of patient data results in a marked improvement in performance over patient data only. Finally, we evaluate how FROC performance on patient test data (EMBED) is affected by the addition of synthetic

examples (from T-SYNTH) and find that addition of T-SYNTH data improves performance at all levels of the image FPR, particularly for scattered and dense breast examples (see Appendix).

5.4 Comparisons to a Generative AI Model

We also compare the performance of T-SYNTH to stable diffusionSDXL (Podell et al., 2023), which generates images of 1024x1024 resolution. We generate images using this model with and without finetuning, where finetuning is done using the textual inversion (Gal et al., 2022) personalization method using a subset of 16 lesions from the patient dataset. We report results using the following controlled sets of training data: T-SYNTH, EMBED baseline, Subset EMBED baseline, Diffusion, and Finetuned Diffusion (see Appendix).

At a high level, the idea is to evaluate whether T-SYNTH and diffusion can supplement positive cases (with lesions present) which are underrepresented in the EM-BED dataset. Our experiments indicate that both types of synthetic images (diffusion and T-SYNTH) improve performance over the under-sampled patient baseline. In the overall performance comparison (see Fig. 6a), the diffusion model without finetuning performs worse than T-SYNTH while the finetuned diffusion model performs better than T-SYNTH. Note that this is expected since T-SYNTH distributions have not been matched to patient samples in EMBED. Similar comparative performance trends are observed across subgroups (see Fig. 6b), although the magnitude of improvement is different across the four considered breast densities.

6. Limitations and Future Work

A number of limitations exist with T-SYNTH as well as our analysis, which can be improved in future iterations. First, there is a domain gap between T-SYNTH and patient DM and DBT images. Much of this gap is due to differences in image reconstruction and postprocessing algorithms, which can greatly impact background intensity, picture sharpness, and tissue contrasts. It is important to mitigate these differences as best as possible, although this is challenging because commercial algorithms that are proprietary. Shifts can be addressed either by adjusting KB parameters, or domain shift mitigation techniques (Billot et al., 2023; Liu et al., 2022). There is also room for improvement in the lesion model used in T-SYNTH, which does not currently capture. We followed a similar generation methodology as that of M-SYNTH dataset (Sizikova et al., 2022) to enable a direct comparison between c-view and mammography, which can be further improved.

In addition, patient datasets, including EMBED, feature complex annotations of findings that are not limited to just

(a) Overall Performance
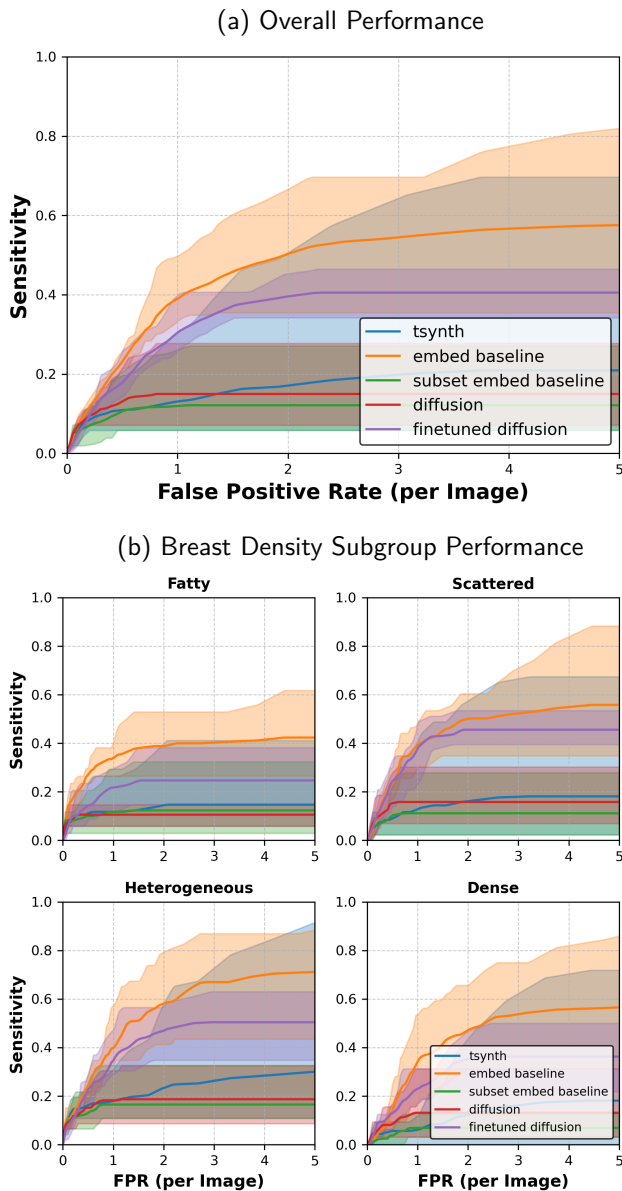

(b) Breast Density Subgroup Performance

Figure 6: Comparison of T-SYNTH, diffusion and finetuned diffusion synthetic data supplementation.

masses. For example, classifications and focal asymmetries are noted, and these findings are often classified by their suspected severity via the BIRADs scale. Thus, our model can be expanded in future work to include other important abnormalities such as calcifications, as well as an improved lesion model that captures diversity of both expression and cancer severity that is seen in patient datasets.

Finally, future work can also expand our analysis to the 3D DBT images in T-SYNTH. Currently, a robust public dataset of 3D DBT images for comparison, with rich annotations such as breast density classifications, is lacking (further highlighting the need for synthetic data). Additionally, while our preliminary findings noted improvement in detection performance on patient data when training data was augmented, there is substantial variation in results

between test trials. This is likely due to the relatively small test and validation sets, which are limited by the relatively few number of fatty and dense breast samples containing findings.

Acknowledgments

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

The authors have no conflict of interests to declare.

Data Availability

Data/Code Location All data and pretrained models are available on Hugging Face [1]. We also uploaded all pretrained models along with their descriptions. Using these models, it is possible to post-process and reproduce all images used in this paper. The code used to train the models available on GitHub in T-SYNTH repository[2].

Potential Use Cases The published data can be used to train detection or segmentation models in combination with patient data. As shown in the manuscript, when patient data is limited, the training can benefit from inclusion of synthetic data, especially for underrepresented subgroups.

Licensing All data and code are released under the CC0-1.0 license. This permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Pretrained models are also shared under the same license.

Ethical Considerations This work is using synthetically generated data and does not include any patient-identifiable information. Publication of synthetic data aims to promote transparency, reproducibility, and fairness in medical AI research.

---

1. https://huggingface.co/datasets/didsr/tsynth
2. https://github.com/DIDSR/tsynth-release

7. Appendix

7.1 Additional T-SYNTH Visualizations

In Fig. 7, we include examples of T-SYNTH images with varying lesion densities. In Fig. 8, we include examples of T-SYNTH images with varying lesion sizes.
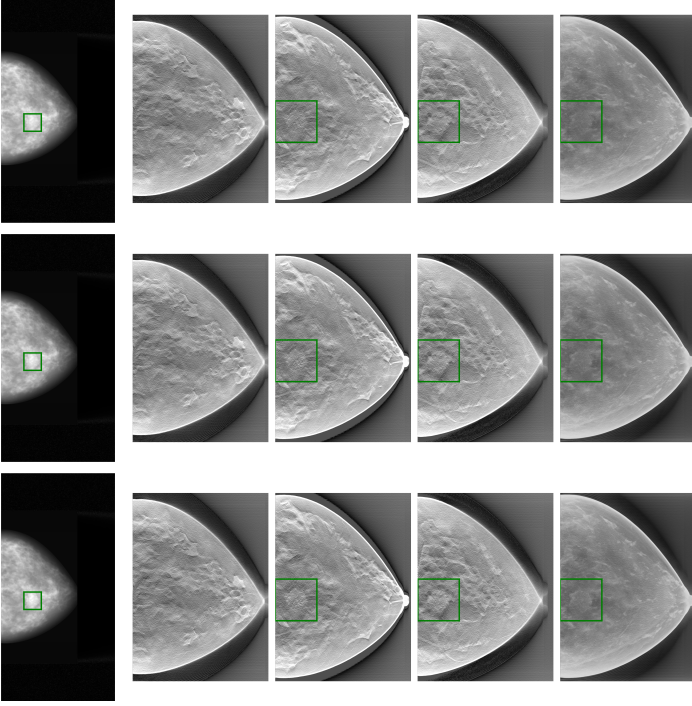

Figure 7: Example T-SYNTH images with varying lesion densities, as measured relative to fibroglandular density (row 1: 1.0, row 2: 1.06, row 3: 1.1). Column 1: DM, columns 2-4: DBT slices at quarter, half, and three-quarters of the total volume depth, column 5: DBT C-View.
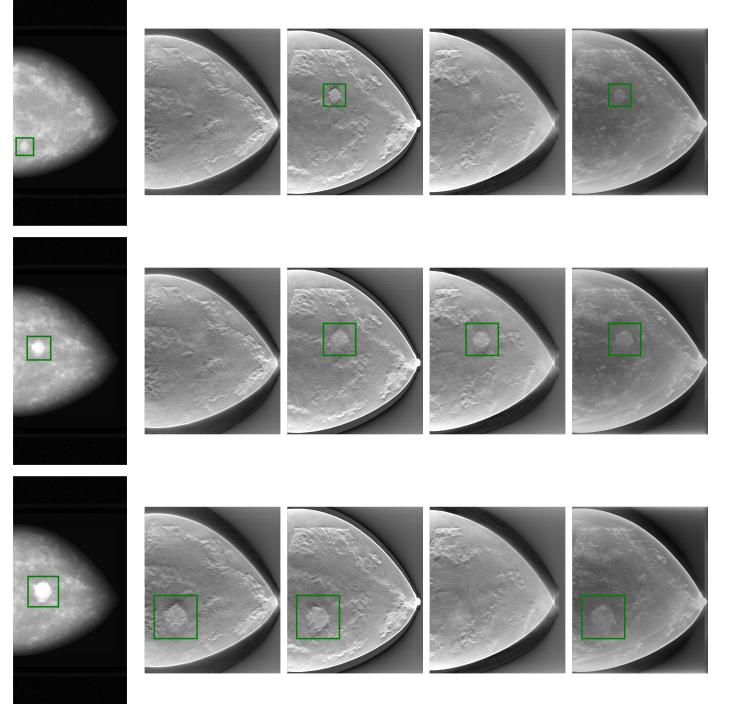

Figure 8: Example T-SYNTH images with varying lesion sizes (row 1: 5mm, row 2: 7mm, row 3: 9mm). Column 1: DM, columns 2-4: DBT slices at quarter, half, and three-quarters of the total volume depth, column 5: DBT C-View.

7.2 Dataset Splits

In Tab. 2 and Tab. 3, we provide a distribution of image and patient samples from the EMBED dataset. In Tab. 4, we provide a distribution of patient samples from the T-SYNTH dataset (the corresponding image samples table can be found in the main paper). The difference across patient and images samples is that the same patient may contain multiple images (e.g., from different time points).

Breast Density	Train Set	Val Set	Test Set	Total
Fatty	53/8614	34/34	29/29	116/8977
Scattered	630/33812	35/35	38/38	703/33885
Hetero	660/33391	34/34	38/38	732/33463
Dense	33/4888	34/34	31/31	98/4953
Total	**1346/80705**	**137/137**	**136/136**	**1649/81278**

Table 2: Distribution of **image** samples ($+/-$) from the EMBED dataset (DM and DBT).

Breast Density	Train Set	Val Set	Test Set	Total
Fatty	-	-	-	-
Scattered	-	-	-	-
Hetero	-	-	-	-
Dense	-	-	-	-
Total	**723/8334**	**73/0**	**78/0**	**874/8334**

Table 3: Distribution of **patient** samples ($+/-$) from the EMBED dataset (DM and DBT).

Breast Density	Train Set	Val Set	Test Set	Total
Fatty	100/100	25/25	25/25	150/150
Scattered	100/100	25/25	25/25	150/150
Hetero	100/100	25/25	25/25	150/150
Dense	100/100	25/25	25/25	150/150
Total	**400/400**	**100/100**	**100/100**	**500/500**

Table 4: Distribution of **patient** samples ($+/-$) from the T-SYNTH dataset (DM and DBT). Examples with same phantom but different lesion size or lesion density are treated as the same patient.

7.3 Additional FROC Analysis on Data Replacement

Finally, in Fig. 9 we evaluate the effect of different patient-to-synthetic training data ratios on detection performance, measured by the area under FROC (FROC-AUC). For each group, green bars represent models trained on partial patient training data relative to the full patient training dataset (800 images, consisting of 400 examples with lesions and 400 examples without lesions). Orange bars represent models trained on a mix of patient and synthetic training data which also has 800 images, but contains proportions (patient:synthetic) of patient and synthetic data. Blue bars show performance of models trained on the full patient dataset. We find that replacing nearly 60% of patient data with synthetic examples achieves the same average level of performance as using the full patient training dataset.
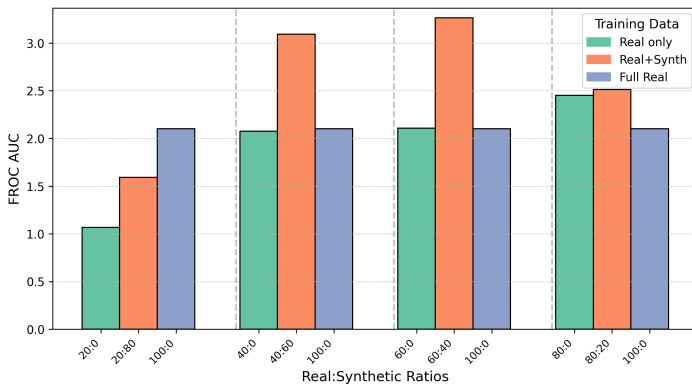

Figure 9: Comparison of FROC AUC scores across different real-to-synthetic training data ratios. Each bar represents the average performance across five runs. The blue bar - 100% patient data is included in each group for reference. The green bar is different proportion of patient data and orange is replacement in different proportion with T-SYNTH data.

7.4 Comparison of DM and DBT

If we compare detection results on DM and DBT (see Fig. 10), we find that the corresponding results from DM, while having slightly lower and more variant performance, are generally consistent with the trends in DBT. Synthetic data appears to generate a slightly larger performance improvement in DBT compared to DM (e.g., compare scattered and dense subgroups).

7.5 Table of Public Breast Image Datasets

Finally, we provide a summary of existing public breast image datasets (both patient and synthetic) in Tab. 5 for additional information.

7.6 Diffusion Experiments

The diffusion model is run using an auto-inpainting pipeline from Hugging Face, where an image without a lesion, a circular mask, and a prompt ("draw a lesion") are presented to the model, which outputs a breast image with the lesion drawn in the specified location. This approach allowed us to generate both images and bounding boxes for each generated lesion.

- **T-SYNTH**: positive (300 from T-SYNTH + 100 from EMBED), negative (400 from EMBED)

- **EMBED baseline**: positive (400 from EMBED), negative (400 from EMBED)

- **Subset EMBED baseline**: positive (100 from EMBED), negative (400 from EMBED)

- **Diffusion**: positive (300 from un-finetuned diffusion, 100 from EMBED), negative (400 from EMBED)

- **Finetuned Diffusion**: positive (300 from finetuned diffusion, 100 from EMBED), negative (400 from EMBED)
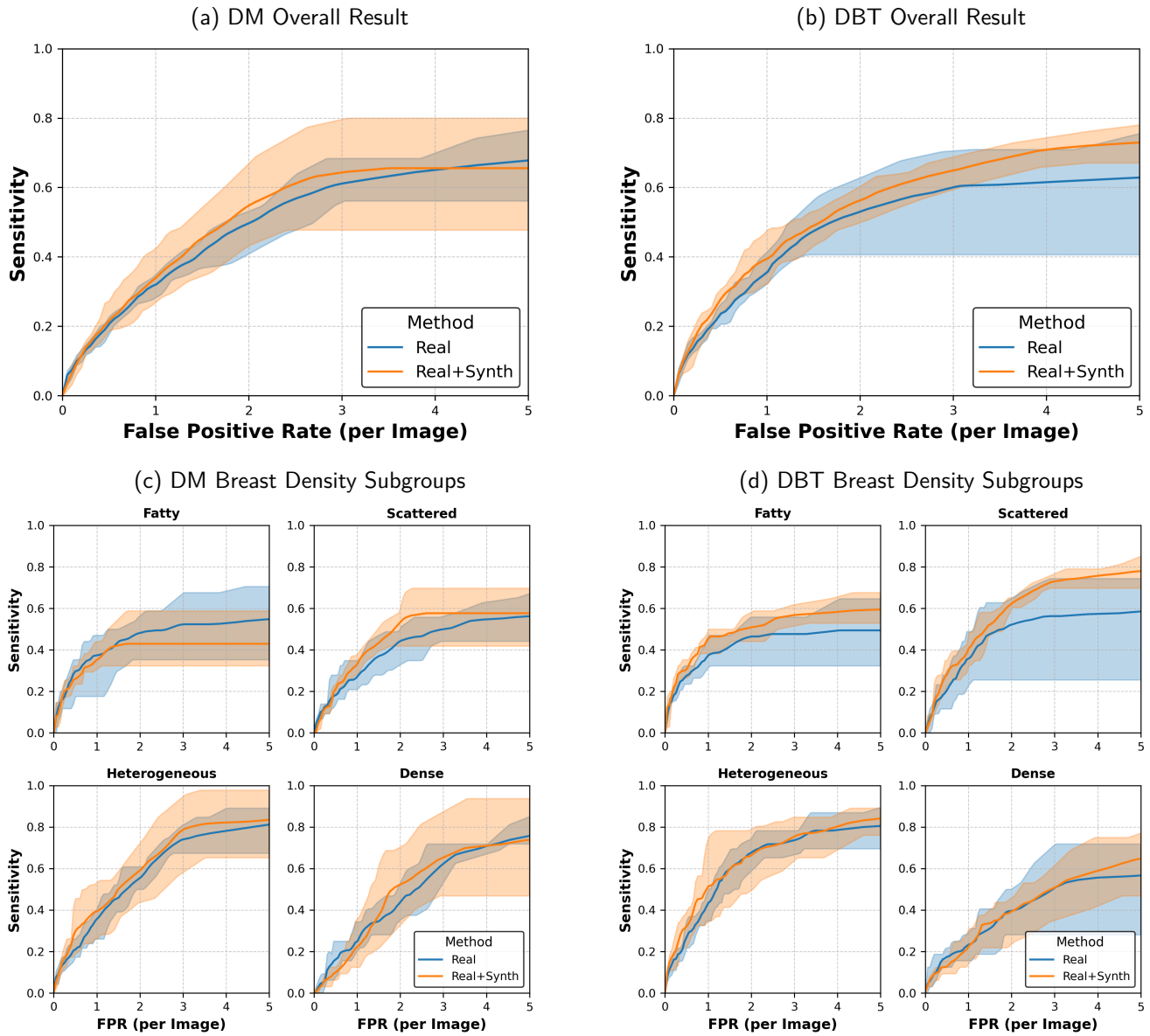
Figure 10: Comparison of Digital Mammography (DM) and Digital Breast Tomosynthesis (DBT) Free-response ROC curve on real test set for detection models trained with real (EMBED Jeong et al. (2023)) and synthetic (T-SYNTH) data. Solid blue and orange lines represent average sensitivity across trials; spread captures minimum and maximum sensitivity.

Patient datasets									
Dataset	DM	DBT	Bi-RADS	Density	# cases	# images	Image categories	Pop	Seg. Annotation
BCS-DBT (Buda et al., 2020)	No	Yes	Yes (5)	No	5,060	22,032	Cancer benign actionable normal	USA	336 mass bboxes 99 arch.distortions bboxes
ADMANI (Frazer et al., 2022)	Yes	No	No	No	629,863	4,411,263[a]	Normal recall	Several nations	Annotations provided **where available**
EMBED (Jeong et al., 2023)	Yes	Yes	Yes (6)	Yes (4)[m]	116,000	3,383,659[b]	Invasive cancer non-invasive cancer high risk	USA[c]	40,000 lesion-level ROIs with annotations [h]
CMMD (Cui et al., 2021)	Yes	No	No	No	1,775	3,712	Benign, malignant	China	None
INBreast (Moreira et al., 2012)	Yes	No	Yes	Yes (3)	115	410	Benign, malignant, normal	Portugal	105 annotated images
OPTIMAM (Halling-Brown et al., 2020)	Yes	No	No	No	172,282	3,072,878	Normal, interval cancers, benign, malignant	UK	7,143 bboxes of lesions
VinDR (Nguyen et al., 2022)	Yes	No	Yes (6)	Yes (4) [z]	5,000	20,000	Normal, interval cancers, benign, malignant	UK	20,000 bboxes of lesions
CBIS-DDSM (Lee et al., 2017)	Yes	No	Yes (5)	Yes (4) [z]	6,775	10,239	normal, benign, malignant	USA	1,644 ROIs of masses [h]
BCDR (Lopez et al., 2012)	Yes	No	Yes (5)	Yes (3) [l]	1,010	3,703	masses microcalcifications calcifications stromal distortions architectural distortions axillary adenopathy	USA	2,335 lesion segmentations
DMID (Oza et al., 2023)	Yes	No	Yes (6)	Yes (3) [l]	510	510	mass-benign mass-malignant calcifications architectural distortion asymmetry	India	510 pixel-level annotations (masks) of masses
CSAW (Dembrower et al., 2019)	Yes	No	No	No	499,807	≈2,000,000	cancerous, healthy	Sweden	1891 tumor pixel-level annotations
KAU-BCMD (Alsolami et al., 2021)	Yes	No	Yes (5)	Yes (4) [z]	1,416	5,662	normal, benign, and malignant	Saudi Arabia	segmentations and bboxes for 5,662 masses
Sheba (Zamir et al., 2021)	Yes	No	No	No	19	42	isolated, clustered microcalcifications	Korea	42 masks of microcalcifications
UC-Davis (Sarno et al., 2021a)	No	Yes	No	No	150	150	N/A	USA	150 pixel-level lesion annotations[j]
MIAS Mammography (Strand, 2022)	Yes	No	No	Yes (3)[l]	416	416	normal, benign, malignant	USA	416 ROIs of masses [i]
NYU [Not publicly available]	Yes	No	Yes (3)	Yes (5) [z]	229,426	1,001,093	cancerous, healthy	USA	8,080 pixel-level mass segmentations
Synthetic datasets									
Dataset	DM	DBT	Bi-RADS	Density	# cases	# images	Image categories	Pop	Seg. Annotation
Sarno (Sarno et al., 2021b)	Yes	Yes	–	–	–	150[d]	Normal	–	No
VICTRE (Badano et al., 2018)	Yes	Yes	–	–	–	2986	Negative, positive cohort	–	Yes[e]
CSAW (Pinaya et al., 2023)	Yes	No	–	–	–	100k	high, medium, low masking levels	–	Yes[n]
M-SYNTH (Sizikova et al., 2024)	Yes	No[g]	–	–	–	45,000	Negative, positive cohort	–	Yes[f]
T-SYNTH (Ours)	Yes	Yes[g]	–	–	–	9,000	Negative, positive cohort	–	Yes[f]

[a] subset available for the RSNA Cancer Detection AI challenge.
[b] 20% available via AWS; contains annotated lesions.
[c] equal representation of African American and White
[d] 150 uncompressed, 60 compressed images
[e] four breast densities, same lesions across all positive cohort
[f] 3 lesion densities, 3 lesion sizes, 4 breast densities, 5 different doses
[g] A corresponding DBT image dataset will be provided in a future release of the dataset.
[h] location of centroid + radius. Saved as a white circle directly onto a copy of the original mammogram, generating a screen save image.
[i] location of centroid + radius.
[j] contains computational digital breast phantoms. Each image voxel was classified in one out of the four main materials presented in the field of view: fibroglandular tissue, adipose tissue, skin tissue, and air
[l] fatty, fatty-glandular, dense-glandular
[m] high density, isodense, low density, fat containing
[n] diffusion-based generator
[z] almost entirely fatty (1), scattered areas of fibroglandular density (2), heterogeneously dense (3), extremely dense (4), unknown (5)

Table 5: Summary of existing breast imaging datasets. The proposed T-SYNTH dataset is the largest synthetic paired DM-DBT dataset available.

References

Asmaa S Alsolami, Wafaa Shalash, Wafaa Alsaggaf, Sawsan Ashoor, Haneen Refaat, and Mohammed Elmogy. King abdulaziz university breast cancer mammogram dataset (KAU-BCMD). *Data*, 2021.

Akram M Asbeutah, Nouralhuda Karmani, AbdulAziz A Asbeutah, Yasmin A Echreshzadeh, Abdullah A AlMajran, and Khalid H Al-Khalifah. Comparison of digital breast tomosynthesis and digital mammography for detection of breast cancer in kuwaiti women. *Med Princ Pract*, 2019.

Andreu Badal and Aldo Badano. Accelerating monte carlo simulations of photon transport in a voxelized geometry using a massively parallel graphics processing unit. *Medical Physics*, 2009.

Aldo Badano, Christian G Graff, Andreu Badal, Diksha Sharma, Rongping Zeng, Frank W Samuelson, Stephen J Glick, and Kyle J Myers. Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial. *JAMA Network Open*, 11 2018.

Aldo Badano, Migue Lago, Elena Sizikova, Jana G. Delfino, Shuyue Guan, Mark A Anastasio, and Berkman Sahiner. The stochastic digital human is now enrolling for in silico imaging trials–methods and tools for generating digital cohorts. *arXiv preprint arXiv:2301.08719*, 2023.

Jun Bai, Russell Posner, Tianyu Wang, Clifford Yang, and Sheida Nabavi. Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: A review. *Medical Image Analysis*, 2021.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*, 2023.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and

Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 2024.

Mateusz Buda, Ashirbani Saha, Ruth Walsh, Sujata Ghate, Nianyi Li, Albert Święcicki, Joseph Y Lo, and Maciej A Mazurowski. Detection of masses and architectural distortions in digital breast tomosynthesis: a publicly available dataset of 5,060 patients and a deep learning model. *arXiv preprint arXiv:2011.07995*, 2020.

Chunyan Cui, Li Li, Hongmin Cai, Zhihao Fan, Ling Zhang, Tingting Dan, Jiao Li, and Jinghua Wang. The chinese mammography database (CMMD): An online mammography database with biopsy confirmed types for machine diagnosis of breast. *Data Cancer Imaging Arch*, 2021.

Karin Dembrower, Peter Lindholm, and Fredrik Strand. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the cohort of screen-aged women (CSAW). *Journal of Digital Imaging*, 2019.

Carol E DeSantis, Stacey A Fedewa, Ann Goding Sauer, Joan L Kramer, Robert A Smith, and Ahmedin Jemal. Breast cancer statistics, 2015: Convergence of incidence rates between black and white women. *CA: a cancer journal for clinicians*, 2016.

Helen ML Frazer, Jennifer SN Tang, Michael S Elliott, Katrina M Kunicki, Brendan Hill, Ravishankar Karthik, Chun Fung Kwok, Carlos A Peña-Solorzano, Yuanhong Chen, Chong Wang, et al. ADMANI: Annotated digital mammograms and associated non-image datasets. *Radiology: Artificial Intelligence*, 2022.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Christian G Graff. A new, open-source, multi-modality digital breast phantom. In *Medical Imaging 2016: Physics of Medical Imaging*. SPIE, 2016.

Mark D. Halling-Brown, Lucy M. Warren, Dominic Ward, Emma Lewis, Alistair Mackenzie, Matthew G. Wallis, Louise S. Wilkinson, Rosalind M. Given-Wilson, Rita McAvinchey, and Kenneth C. Young. OPTIMAM mammography image database: A large-scale resource of mammography images and clinical data. *Radiology: Artificial Intelligence*, 2020.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis. (ICCV)*, 2017.

Jiwoong J Jeong, Brianna L Vey, Ananth Bhimireddy, Thomas Kim, Thiago Santos, Ramon Correa, Raman Dutt, Marina Mosunjac, Gabriela Oprea-Ilies, Geoffrey Smith, et al. The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiology: Artificial Intelligence*, 2023.

Varun A. Kelkar, Dimitrios S. Gotsis, Frank J. Brooks, Prabhat KC, Kyle J. Myers, Rongping Zeng, and Mark A. Anastasio. Assessing the ability of generative adversarial networks to learn canonical medical image statistics. *IEEE Transactions on Medical Imaging*, 2023.

DS Klein, MA Lago, CK Abbey, and MP Eckstein. A 2D synthesized image improves the 3D search for foveated visual systems. *IEEE Transactions on Medical Imaging*, Aug 2023.

Nicholas Konz, Mateusz Buda, Hanxue Gu, Ashirbani Saha, Jichen Yang, Jakub Chłędowski, Jungkyu Park, Jan Witowski, Krzysztof J Geras, Yoel Shoshan, et al. A competition, benchmark, code, and data for using artificial intelligence to detect lesions in digital breast tomosynthesis. *JAMA Network Open*, 2023.

Prudence B. Lam, Pamela M. Vacek, Berta M. Geller, and Hyman B. Muss. The association of increased weight, body mass index, and tissue density with the risk of breast carcinoma in vermont. *Cancer*, 2000.

Rebecca Lee, Francisco Gimenez, Assaf Hoogi, Kanae Miyake, Mia Gorovoy, and Daniel Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 12 2017.

Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, Jonghye Woo, et al. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, 2022.

Miguel G Lopez, Naimy Posada, Daniel C Moura, Raúl Ramos Pollán, José M Franco Valiente, César Suárez Ortega, Manuel Solar, Guillermo Diaz-Herrero, IMAP Ramos, Joana Loureiro, et al. BCDR: a breast cancer digital repository. In *International Conference on Experimental Mechanics*, 2012.

Ana Luiza Menegatti Pavan, Marcela de Oliveira, Matheus Alvarez, Ana Júlia Martins Sampaio, Andre Petean Trindade, Sergio Barbosa Duarte, and Diana Rodrigues

de Pina. Breast tissue segmentation by fuzzy c-means. *Physica Medica*, 2016.

Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic Radiology*, 2012.

Hieu T. Nguyen, Ha Q. Nguyen, Hieu H. Pham, Khanh Lam, Linh T. Le, Minh Dao, and Van Vu. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *medRxiv*, 2022.

Parita Rajiv Oza, Urvi Oza, Rajiv Oza, Paawan Sharma, Samir Patel, Pankaj Kumar, and Bakul Gohel. Digital mammography dataset for breast cancer diagnosis research (DMID) with breast mass segmentation analysis. *Biomedical Engineering Letters*, 2023.

Walter H. L. Pinaya, Mark S. Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F. da Costa, Ashay Patel, Hyungjin Chung, Can Zhao, Wei Peng, Zelong Liu, Xueyan Mei, Oeslle Lucena, Jong Chul Ye, Sotirios A. Tsaftaris, Prerna Dogra, Andrew Feng, Marc Modat, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Generative AI for medical imaging: extending the MONAI framework, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

José Luis Raya-Povedano, Sara Romero-Martín, Esperanza Elías-Cabot, Albert Gubern-Mérida, Alejandro Rodríguez-Ruiz, and Marina Álvarez-Benito. Ai-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation. *Radiology*, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015.

Wessam M. Salama and Moustafa H. Aly. Deep learning in mammography images segmentation and classification: Automated cnn approach. *Alexandria Engineering Journal*, 2021.

Antonio Sarno, G. Mettivier, Francesca di Franco, Antonio Varallo, Kristina Bliznakova, Andrew Hernandez, John Boone, and Paolo Russo. Dataset of patient-derived digital breast phantoms for in-silico studies in breast computed tomography, digital breast tomosynthesis and digital mammography. *Medical Physics*, 03 2021a.

Antonio Sarno, Giovanni Mettivier, Francesca di Franco, Antonio Varallo, Kristina Bliznakova, Andrew M. Hernandez, John M. Boone, and Paolo Russo. Dataset of patient-derived digital breast phantoms for in silico studies in breast computed tomography, digital breast tomosynthesis, and digital mammography. *Medical Physics*, 2021b.

Aunnasha Sengupta, Miguel A Lago, and Aldo Badano. In situ tumor model for longitudinal in silico imaging trials. *Physics in Medicine & Biology*, 2024.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, jul 2024.

Elena Sizikova, Xu Cao, Ashia Lewis, Kenny Moise, and Megan Coffee. Improving computed tomography (CT) reconstruction via 3D shape induction. *arXiv preprint arXiv:2208.10937*, 2022.

Elena Sizikova, Niloufar Saharkhiz, Diksha Sharma, Miguel Lago, Berkman Sahiner, Jana Delfino, and Aldo Badano. Knowledge-based in silico models and dataset for the comparative evaluation of mammography AI for a range of breast characteristics, lesion conspicuities and doses. *Advances in Neural Information Processing Systems*, 2024.

Brian L Sprague, Rebecca Yates Coley, Kathryn P Lowry, Karla Kerlikowske, Louise M Henderson, Yu-Ru Su, Christoph I Lee, Tracy Onega, Erin JA Bowles, Sally D Herschorn, Roberta M. diFlorio Alexander, and Diana L. Miglioretti. Digital breast tomosynthesis versus digital mammography screening performance on successive screening rounds from the breast cancer surveillance consortium. *Radiology*, June 2023.

F Strand. CSAW-CC (mammography)—a dataset for AI research to improve screening, diagnostics and prognostics of breast cancer, 2022.

Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for

36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 2021.

Volkan M. Tiryaki and Veysel Kaplanoğlu. Deep learning-based multi-label tissue segmentation and density assessment from mammograms. *IRBM*, 2022.

Mika Yamamuro, Yoshiyuki Asai, Naomi Hashimoto, Nao Yasuda, Hiorto Kimura, Takahiro Yamada, Mitsutaka Nemoto, Yuichi Kimura, Hisashi Handa, Hisashi Yoshida, Koji Abe, Masahiro Tada, Hitoshi Habe, Takashi Nagaoka, Seiun Nin, Kazunari Ishii, and Yohan Kondo. Utility of u-net for the objective segmentation of the fibroglandular tissue region on clinical digital mammograms. *Biomedical Physics & Engineering Express*, jun 2022.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, Atul J Butte, and Emily Alsentzer. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, Jan 2024.

Roee Zamir, Shai Bagon, David Samocha, Yael Yagil, Ronen Basri, Miri Sklair-Levy, and Meirav Galun. Segmenting microcalcifications in mammograms and its applications. *CoRR*, 2021.

Jiadong Zhang, Jiaojiao Wu, Xiang Sean Zhou, Feng Shi, and Dinggang Shen. Recent advancements in artificial intelligence for breast cancer: Image augmentation, segmentation, diagnosis, and prognosis approaches. *Seminars in Cancer Biology*, 2023.

Wenlian Zhu, Susan Harvey, Katarzyna J. Macura, David M. Euhus, and Dmitri Artemov. Invasive breast cancer preferably and predominantly occurs at the interface between fibroglandular and adipose tissue. *Clinical Breast Cancer*, 2017.