Knowledge-based in silico models and dataset for the comparative evaluation of mammography AI

E. Sizikova, N. Saharkhiz, D. Sharma, M. Lago, B. Sahiner, J. G. Delfino, A. Badano
Office of Science and Engineering Laboratories
Center for Devices and Radiological Health
U.S. Food and Drug Administration
Silver Spring, MD 20993 USA

Abstract

To generate evidence regarding the safety and efficacy of artificial intelligence (AI) enabled medical devices, AI models need to be evaluated on a diverse population of patient cases, some of which may not be readily available. We propose an evaluation approach for testing medical imaging AI models that relies on in silico imaging pipelines in which stochastic digital models of human anatomy (in object space) with and without pathology are imaged using a digital replica imaging acquisition system to generate realistic synthetic image datasets. Here, we release M-SYNTH*, a dataset of cohorts with four breast fibroglandular density distributions imaged at different exposure levels using Monte Carlo x-ray simulations with the publicly available Virtual Imaging Clinical Trial for Regulatory Evaluation (VICTRE) toolkit. We utilize the synthetic dataset to analyze AI model performance and find that model performance decreases with increasing breast density and increases with higher mass density, as expected. As exposure levels decrease, AI model performance drops with the highest performance achieved at exposure levels lower than the nominal recommended dose for the breast type.

1 Introduction

The goal of this work is to demonstrate that AI models for medical imaging can be evaluated using simulations, specifically, using an in silico (also known as synthetic) imaging pipeline equipped with a stochastic model for human anatomy and disease [1]. We show that in silico methods can constitute rich sources of data with realistic physical variability for performing comparative analysis of AI device performance.

To date, computational models have been applied to some extent for the analysis of nearly all medical imaging modalities and for a wide variety of clinical tasks [2]. Since it is critical to ensure patient safety and system effectiveness in healthcare applications, rigorous and thorough testing procedures must be performed in order to study performance in the intended population, including subpopulations of interest. To prevent estimates that might be biased by overfitting, model testing is typically performed on a previously unseen dataset. However, datasets consisting of patient images may present a limited distribution of the variability in human anatomy and may not always capture rare, but life-critical cases, and may be biased towards specific populations or parameters of image acquisition devices dominant at specific clinical sites. In addition, patient data and associated health records may not be available due to patient privacy, cost, or additional risk associated with additional imaging procedures. Precise mass location and extent (e.g., mass boundaries) are typically not available in the patient's records, and it is burdensome, error-prone, and sometimes impossible to collect this information retrospectively. In many medical imaging applications, these limitations pose

*Code and data links available at: https://github.com/DIDSR/msynth-release/

NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI.

a significant barrier to development and evaluation of novel computational techniques in medical imaging products.

We propose evaluating AI models using physics-based simulations. We create realistic test cases by imaging digital objects using digital image acquisition systems. Our in silico testing pipeline offers the ability to control both object and acquisition parameters, and generate highly realistic test cases.We show that digital objects and computer-simulated replicas of image acquisition devices offer a rich source of realistic data capturing a variety of patient and imaging conditions for evaluation purposes. In particular, our approach (and associated dataset) allows for performing *comparative* analysis of AI performance across physical breast properties (e.g., mass size) and imaging characteristics (e.g., radiation dose). Such testing typically cannot be performed with patient data, as the data may be too costly to collect or unsafe to acquire (e.g., one cannot ethically re-image the same patient multiple times using ionizing radiation). Our contributions in this work can be summarized as follows:

- We demonstrate that, using this approach, we can detect differences in AI model performance based on selected image acquisition device or physical object model parameters. Specifically, we evaluate the effect of image acquisition (radiation dose) and object model (breast and mass densities, mass size) parameters on the performance of the AI model.
- We release a dataset, M-SYNTH, to facilitate testing with pre-computed data using the proposed pipeline. The dataset consists of 1,200 stochastic knowledge-based models and their associated digital mammography (DM) images with varying physical (breast density, mass size and density) and imaging (dose) characteristics.

2 Background

First, we introduce the concepts of knowledge-based models and physics-based imaging simulation that form the *in silico imaging pipeline*, the foundation of our work.

Object Models. Knowledge-based (KB) models incorporate information about the physical world into the data generation process to create realistic virtual representations of human parts or organs [3]. As discussed in [1], large cohorts of digital stochastic human models can be represented by:

$$\{\mathbf{f}_s\}_{s=1}^{S} = \sum_n \theta_n^s \phi_n(\boldsymbol{r}), \tag{1}$$

where s denotes a particular state or random realization of a digital human in a cohort of size S, $\boldsymbol{r}$ denotes a spatial variable, ϕ_n denote expansion (basis) functions, and θ_n denote expansion coefficients. Knowledge-based models are specifically constructed by sampling a set of θ_n in Eq. 1 from distributions representing the relevant model characteristics, given a specific ϕ_n based on the application. The characteristics of the distributions are often derived from physical or biological measurements. In the case of breast, knowledge-based models allow us to vary physical patient characteristics, including breast size, breast shape, mass size, and mass density. Specifically, the object (breast) is a model D, parameterized by a vector x characterizing a fixed, user-defined set of physiological properties (e.g., breast density, mass presence, mass size, glandularity). Given a sample x_s, we can generate a realistic, high-resolution object $f_s = D(x_s)$. We rely on Graff's breast model [3] as the KB model for this project and describe its properties in Section 3.

Digital Mammography (DM) image generation. Once created, KB models are imaged using simulations of x-ray transport through the materials present in each KB model. The image acquisition device I is a parametric model that receives the object d_i as well as user-defined choices for control parameters y (e.g., detector type, radiation dose) and outputs an image $r_{i,j} = I(d_i, y_j)$ given a sample choice of parameters y_j and an input object d_i. Parameters of such a system (e.g., geometry, source characteristics, detector technology, anti-scatter grid, etc.) can emulate system geometries and x-ray acquisition parameters found in commercially available imaging device (e.g., mammography) specifications. In our work, we used MC-GPU [4], a Monte Carlo x-ray simulation software implemented on GPUs that generates mammography images. Additional details for this component of the pipeline can be found in Section 3.

3 Dataset Generation

The use of in silico imaging allows for the generation of large object and image datasets without the need for human clinical trials. Here, we take advantage of the benefits of the in silico approach to perform comparative analysis of AI model performance across different physical properties of the case population of breast models. We rely on the VICTRE pipeline [†] for generating breast models and their corresponding DM images. Previous work [5] has shown that the VICTRE pipeline replicated the results of a clinical study comparing DM and digital breast tomosynthesis (DBT) involving hundreds of enrolled women.

Breast Model Synthesis. In silico breast models [3] (also known as breast imaging phantoms) were generated using a procedural analytic model which allows for adjusting various patient characteristics, including breast shape, size and glandular density. The models are compressed in the craniocaudal direction using FEBio [6], an open source finite-element software. We simplified the breast materials into non-glandular (as fat) or glandular tissue with Young's modulus and Poisson ratio of $E = 5Pa$, $\nu = 0.49$ and $E = 15Pa$, $\nu = 0.49$, respectively. Lesions were inserted in a subset to create the signal-present cohort. These models were then imaged using a state-of-the-art Monte Carlo x-ray transport code (MC-GPU) [4]. We studied breast densities of extremely dense (referred to as "dense"), heterogeneously dense (referred to as "hetero"), scattered, and fatty, matching the distributions from [5]. For each breast density, a different breast size is used to correspond with population statistics. Therefore, the dense breast is the smallest, followed by heterogeneously dense, then scattered, and then fatty. Each breast model was compressed to 3.5 cm, 4.5 cm, 5.5 cm, and 6.0 cm for each respective density, mimicking the organ compression during the imaging. Random spiculated breast masses were generated using the de Sisternes model [7] with three different sizes (5 mm, 7 mm and 9 mm radii) and mass density was set to be a factor of glandular tissue density (1.0, 1.06 and 1.1 times). Note that for dense and hetero breasts, we only used mass sizes of 5 and 7 mm, since 9 mm masses do not fit within the breast region. No micro-calcification clusters were inserted. To create the signal-present cohort, a single spiculated mass was inserted in half of the cases at randomly chosen locations chosen from a list of candidate sites determined by the position of the terminal duct lobular units. The resulting in silico dataset comprises of 1,200 digital breast models, corresponding to 300 patients per breast size/density. Compared to the original VICTRE trial [5], we introduce variations in mass size and density.

Digital Mammography (DM) Generation. To simulate the x-ray imaging process, we used MC-GPU [4], a Monte Carlo x-ray simulation software implemented on GPUs that generates DM images. The detector model relies on system geometries and x-ray acquisition parameters inspired by the currently available Siemens Mammomat Inspiration DM system. The dosimetric and x-ray acquisition parameters were selected based on publicly available device specifications and clinical recommendations for each compressed breast thickness and glandularity (20-100% of the clinically recommended dose), see [4, 8] for additional details. The detector model (known as DIR in [8]) is representative of a solid-state amorphous selenium transducer in a direct detector configuration.

4 Results and Analysis

In this section, we present an approach to using our M-SYNTH dataset to evaluate an AI device. Formally, an image processing AI model F takes as input an image r and predicts a specific property of interest $F(r)$ about the image. For example, such a model can predict the presence or absence of a mass. Typically for AI models, F is a neural network and is trained on a dataset of images and their labels $T_{train} = \{(r_1, l_1), (r_2, l_2), \dots (r_n, l_n)\}$, and then evaluated on a held-out dataset T_{test}. When using patient images, evaluation is limited to the variability contained in the samples and in the annotations present across examples in the fixed test set T_{test}. Instead, we propose to generate T_{train} and T_{test} dynamically using D and I described above in order to test F across variations in model x and acquisition parameters y.

4.1 Implementation Details

Evaluation Metrics We evaluate performance using the area under curve (AUC) metric for a mass detection task. Specifically, we treat evaluation as a multiple-reader multiple-case study, where an

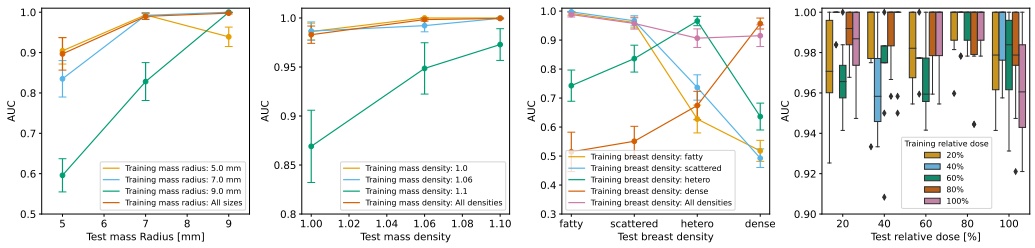
[†] See VICTRE Github Page and FDA Regulatory Science Tools (RST) Catalog.

AI model is a single reader. Multiple readers are obtained by re-training the model with different random seeds. We rely on the iMRMC software [9, 10] to identify associated confidence intervals.

Network Training We represent the AI-enabled device as a neural network with an efficientnet_b0 architecture, receiving an image with one channel and dimensions of 224 by 224, and outputting a binary mass presence label. The network is trained with batch size 64 using binary cross entropy loss (BCE) and optimized using RMSProp optimizer (with learning rate 0.0001). We rely on the timm library [11] and fine-tune the model pre-trained with ImageNet [12]. We also compared performance with alternative architectures (vit_small_patch16_224 and vgg_16), but results were very similar (see supplementary material). For each specific breast density, radiation dose level, and mass size and density, the 300 images in the M-SYNTH dataset were divided into 200 for training, 50 for validation, and 50 for testing. For comparison, we also trained the AI device on 410 patient DM images from the INbreast dataset [13], where images were obtained using MammoNovation Siemens full-field digital mammography system with a solid-state amorphous selenium detector. We use the same pre-processing and training regimes on this dataset and learn a network to predict mass presence. The trained models on the real patient dataset were then tested on 50 examples of the M-SYNTH dataset for each specific breast density, dose level, and mass size and density. The full experimental setup is implemented in Python and C over a cluster with 50 Tesla V100-SXM2 GPUs.

4.2 Experimental Results

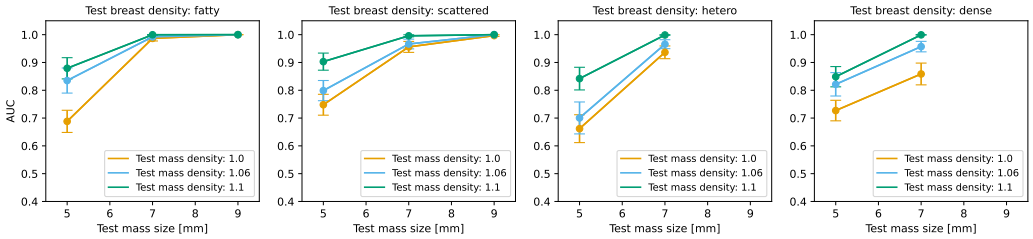
We identify two tasks that can be performed using our method. In the *subgroup analysis* task, we train and test an AI model using the released synthetic (M-SYNTH) dataset to identify performance changes on specified subgroups. In the *patient data evaluation* task, we study how an AI model trained on patient data (INbreast) performs on the proposed M-SYNTH dataset. This task can help identify where the trained model may show variable performance for different subgroups belonging to the target population.



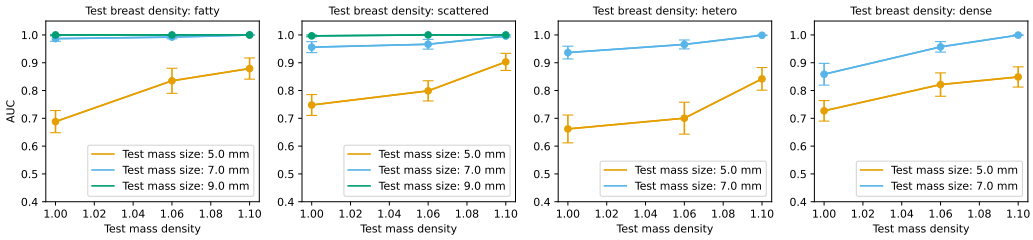
(a) Mass size (b) Mass density (c) Breast density (d) Dose level

Figure 1: *Subgroup analysis*. Performance change across (a) mass size, (b) mass density, (c) breast density, and (d) radiation dose, for models trained and tested on our M-SYNTH dataset. These parameters remained constant for the set of experiments performed during both training and test: (a) Fatty breast phantom, mass density of 1.06, and relative dose of 100%. (b) Fatty breast phantom, mass size of 7 mm, and relative dose of 100%. (c) Mass density of 1.06, mass size of 7 mm, and relative dose of 100%. (d) Fatty breast phantom, mass density of 1.06, and mass size of 7 mm.

Subgroup Analysis. In Figures 1 and 2, we report the results of the AI model performance at detecting masses, when the model is trained and tested on our dataset (see Section 4.1 for details of splits). We find that masses with larger sizes or higher densities (Figures 1a-b) are more easily detected. Although models trained on all sizes or mass densities have the highest performance, when the models are trained on smaller masses or lower densities, they generalize better to other masses (more difficult cases). The performance of the models are highest when they are tested and trained on the same breast density and decrease as the density of the test breast phantom differs from the train phantom (Figures 1c). The dose levels applied in this study have minimal impact on the performance of the models and resulted in similar AUC values (Figures 1d). Evaluation of the performance change across all the breast densities (Figures 2a-b) reveals that the AUC improves with larger mass density and mass size, yet is impacted by the breast density, where mass detection performance is lowest in high-density breasts (dense) and highest in low-density breasts (fatty) in most of the cases, consistent with findings from clinical practice.



(a) AUC as a function of mass size across all breast densities.

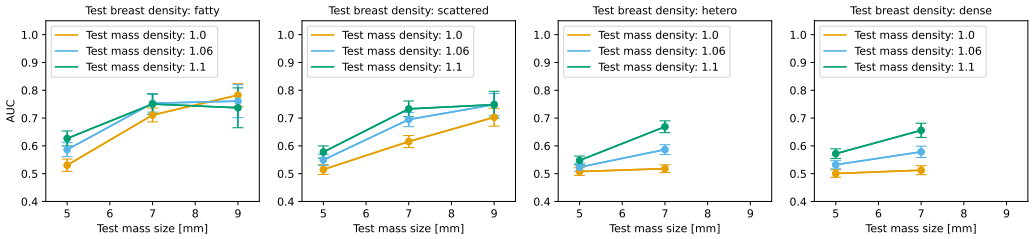


(b) AUC as a function of mass density across all breast densities.

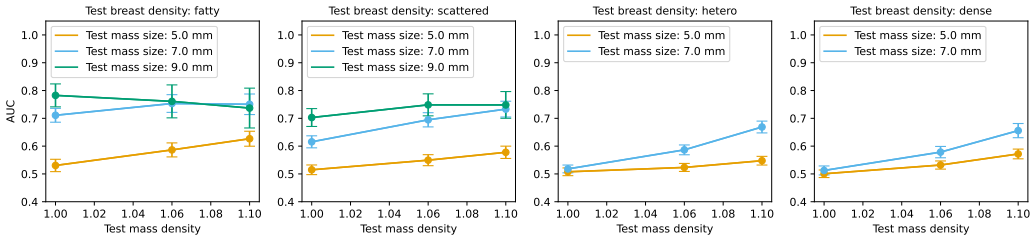
Figure 2: *Subgroup analysis*. Performance changes for models trained and tested on our M-SYNTH dataset. For each data point, the model is trained on 250 images with masses of radii of 7 mm and mass densities of 1.06, and tested on 50 images with mass characteristics shown in plots for each specific breast density. The radiation dose level remains constant at 100% of the clinically recommended dose for each breast density during training and test.

Patient Data Evaluation. In Figure 3, we report experiments where the AI model is trained on INbreast data and evaluated on the M-SYNTH data. Although the performance results for all experiments are lower in general, we find a similar set of trends as when the model is trained on M-SYNTH data. Note that we have made no attempt to match the radiation dose levels or the image acquisition parameters for these comparisons using patient images. Even though the simulated pipeline is designed to replicate a specific DM system with a particular detector technology and technique factors, the comparison suggests similarity between the datasets. The images are qualitatively different but overall have similar glandular patterns which is an important consideration for the realism of the task of detecting masses in a noisy background. We also assessed similarity between the INbreast and M-SYNTH datasets in terms of low-level pixel distributions using first five statistical moments: mean, variance, skewness, kurtosis, and hyperskewness. We found that there is a reasonably good alignment in terms of moments, especially when the synthetic images were included at all four breast densities (see supplementary material). Future work should develop a more detailed comparison, including radiomic features for the training and testing datasets used in the study to complement the validation of our approach.

Limitations. There are a number of limitations to our work. First, simulations may require long runtimes and demand large computational resources, thus somewhat limiting the amounts of data that can be generated. This limitation needs to be considered with respect to the difficulty of obtaining large patient image datasets with known mass locations. In addition, data can be pre-generated offline (as we do with the M-SYNTH dataset), therefore, removing the large runtime limit and computational burden off the user. Second, testing with simulations is constrained to the variability captured by the parameter space of the object models for anatomy and pathology and the acquisition system. Thus, the complexity of the object model and acquisition system may need to be adjusted depending on the complexity of the questions to be investigated with simulated testing. In particular, a potential risk of testing using simulated data is missing the variability observed in patient populations. Finally, there is a risk of mis-judging model performance due to a domain gap between real and synthetic examples. However, the realism and sophistication of object-based modeling of the imaging pipeline are improving rapidly and may soon compete with other approaches, making approaches based on synthetic data useful and practical for regulatory evaluation of AI-enabled medical devices.



(a) AUC as a function of mass size across all breast densities.



(b) AUC as a function of mass density across all breast densities.

Figure 3: *Model Evaluation*. Performance changes for a model trained on 410 real patient images (INbreast dataset) and tested on our M-SYNTH dataset. The test sets consist of 50 images using parameters shown in the plots. The test radiation dose is set to 100% of the clinically recommended dose for each breast density.

5 Conclusion and Future Work

We introduce an approach for validating AI models using physics-based simulations of digital humans from the object space to the image data, specifically for the task of breast cancer mass detection. The simulated images are highly realistic and offer a challenging test case for AI model evaluation. Our findings are consistent with expected performance and show that the AI model performance increases with mass size and mass density as expected. Finally, we show that our approach can be used to validate a model trained on independent patient data. This finding suggests that the proposed simulation setup can be used as a framework for more general evaluation of medical AI devices. The goal of this study is to demonstrate as proof-of-concept the feasibility of using simulated data to evaluate the comparative performance of AI models. In future work, it would be important to assess the evaluation approach for additional parameters in terms of the distribution of the population of digital humans in the object space, and for a range of image acquisition systems (e.g., by considering alternative simulators). By imaging a more diverse population of breast models, we hope to identify additional insights regarding AI evaluation. Finally, it is important to note that the testing is limited to the variability captured in the digital representations and may not fully indicate absolute real-world performance or trends. This study illustrates that physics-based simulation of mammography images can be a least burdensome and cost-efficient approach to evaluate AI model performance for a wide range of scenarios, including a variety of image acquisition parameters and diverse populations that may not be available or are hard to obtain from human studies. Moreover, this approach offers a complementary evaluation paradigm that does not depend on the availability of patient data.

6 Acknowledgements

We thank Andreu Badal (OSEL/CDRH/FDA) and anonymous reviewers for helpful suggestions, Kenny Cha, Mike Mikailov and the OpenHPC team (OSEL/CDRH/FDA) for providing help with experiments, Akhonda, Mohammad (OSEL/CDRH/FDA) for help with data release, and Andrea Kim (OSEL/CDRH/FDA) for rendering visualizations of the 3D breast model. This is a contribution of the US Food and Drug Administration and is not subject to copyright. The mention of commercial products herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

References

[1] A Badano, M Lago, E Sizikova, JG Delfino, S Guan, MA Anastasio, and B 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.

[2] Ana Barragán-Montero, Umair Javaid, Gilmer Valdés, Dan Nguyen, Paul Desbordes, Benoit Macq, Siri Willems, Liesbeth Vandewinckele, Mats Holmström, Fredrik Löfman, et al. Artificial intelligence and machine learning for medical imaging: A technology review. *Physica Medica*, 83:242–256, 2021.

[3] Christian G Graff. A new, open-source, multi-modality digital breast phantom. In *Medical Imaging 2016: Physics of Medical Imaging*, volume 9783, pages 72–81. SPIE, 2016.

[4] Andreu Badal, Diksha Sharma, Christian G Graff, Rongping Zeng, and Aldo Badano. Mammography and breast tomosynthesis simulator for virtual clinical trials. *Computer Physics Communications*, 261:107779, 2021.

[5] 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*, 1(7):e185474–e185474, 11 2018.

[6] Steve A. Maas, Benjamin J. Ellis, Gerard A. Ateshian, and Jeffrey A. Weiss. FEBio: Finite Elements for Biomechanics. *Journal of Biomechanical Engineering*, 134(1):011005, 02 2012.

[7] Luis de Sisternes, Jovan G Brankov, Adam M Zysk, Robert A Schmidt, Robert M Nishikawa, and Miles N Wernick. A computational model to generate simulated three-dimensional breast masses. *Medical physics*, 42(2):1098–1118, 2015.

[8] Aunnasha Sengupta, Andreu Badal, Andrey Makeev, and Aldo Badano. Computational models of direct and indirect x-ray breast imaging detectors for in silico trials. *Medical Physics*, 49(11):6856–6870, 2022.

[9] Brandon D. Gallas, Andriy Bandos, Frank W. Samuelson, and Robert F. Wagner. A framework for random-effects roc analysis: Biases with the bootstrap and other variance estimators. *Communications in Statistics - Theory and Methods*, 38(15):2586–2603, 2009.

[10] RST Catalog. iMRMC: Software for the Statistical Analysis of multi-reader multi-case studies, June 2022.

[11] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[13] 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*, 19(2):236–248, 2012.