Computer-Aided Diagnosis in Multiparametric Magnetic Resonance Imaging Screening of Women With Extremely Dense Breasts to Reduce False-Positive Diagnoses

Erik Verburg, MSc, Carla H. van Gils, PhD,† Marije F. Bakker, PhD,† Max A. Viergever, PhD,* Ruud M. Pijnappel, MD, PhD,‡ Wouter B. Veldhuis, MD, PhD,‡ and Kenneth G. A. Gilhuijs, PhD**

Objectives: To reduce the number of false-positive diagnoses in the screening of women with extremely dense breasts using magnetic resonance imaging (MRI), we aimed to predict which BI-RADS 3 and BI-RADS 4 lesions are benign. For this purpose, we use computer-aided diagnosis (CAD) based on multiparametric assessment.

Materials and Methods: Consecutive data were used from the first screening round of the DENSE (Dense Tissue and Early Breast Neoplasm Screening) trial. In this trial, asymptomatic women with a negative screening mammography and extremely dense breasts were screened using multiparametric MRI. In total, 4783 women, aged 50 to 75 years, enrolled and were screened in 8 participating hospitals between December 2011 and January 2016. In total, 525 lesions in 454 women were given a BI-RADS 3 (n = 202), 4 (n = 304), or 5 score (n = 19). Of these lesions, 444 were benign and 81 were malignant on histologic examination.

The MRI protocol consisted of 5 different MRI sequences: T1-weighted imaging without fat suppression, diffusion-weighted imaging, T1-weighted contrast-enhanced images at high spatial resolution, T1-weighted contrast-enhanced images at high temporal resolution, and T2-weighted imaging. A machine-learning method was developed to predict, without deterioration of sensitivity, which of the BI-RADS 3– and BI-RADS 4–scored lesions are actually benign and could be prevented from being recalled. BI-RADS 5 lesions were only used for training, because the gain in preventing false-positive diagnoses is expected to be low in this group. The CAD consists of 2 stages: feature extraction and lesion classification. Two groups of features were extracted: the first based on all multiparametric sequences, the second based only on sequences that are typically used in abbreviated MRI protocols. In the first group, 49 features were used as candidate predictors: 46 were automatically calculated from the MRI scans, supplemented with 3 clinical features (age, body mass index, and BI-RADS score). In the second group, 36 image features and the same 3 clinical features were used. Each group was considered separately in a machine-learning model to differentiate between benign and malignant lesions. We developed a Ridge regression model using 10-fold cross validation. Performance of the models was analyzed using an accuracy measure curve and receiver-operating characteristic analysis.

Results: Of the total number of BI-RADS 3 and BI-RADS 4 lesions referred to additional MRI or biopsy, 425/487 (87.3%) were false-positive. The full multiparametric model classified 176 (41.5%) and the abbreviated-protocol model classified 111 (26.2%) of the 425 false-positive BI-RADS 3– and BI-RADS 4–scored lesions as benign without missing a malignant lesion.

If the full multiparametric CAD had been used to aid in referral, recall for biopsy or repeat MRI could have been reduced from 425/487 (87.3%) to 311/487 (63.9%) lesions. For the abbreviated protocol, it could have been 376/487 (77.2%).

Conclusions: Dedicated multiparametric CAD of breast MRI for BI-RADS 3 and 4 lesions in screening of women with extremely dense breasts has the potential to reduce false-positive diagnoses and consequently to reduce the number of biopsies without missing cancers.

Key Words: magnetic resonance imaging, machine learning, breast density, screening, false-positives

(*Invest Radiol* 2020;55: 438–444)

Received for publication November 13, 2019; and accepted for publication, after revision, December 21, 2019.

From the *Image Sciences Institute, †Julius Center for Health Sciences and Primary Care, and ‡Department of Radiology, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands.

Conflicts of interest and sources of funding: none declared. This study is financially supported by KWF (grant number UU-2014-7151) and used data acquired during the DENSE trial. The DENSE trial was supported by the regional screening organizations, Volpara Solutions, the Dutch Expert Centre for Screening, and the National Institute for Public Health and the Environment. The DENSE trial is financially supported by the University Medical Center Utrecht (project number: UMCU DENSE), the Netherlands Organization for Health Research and Development (ZonMw, project numbers: ZONMW-200320002-UMCU and ZonMW Preventie 50-53125-98-014), the Dutch Cancer Society (KWF Kankerbestrijding, project numbers: DCS-UU-2009-4348, UU-2014-6859, and UU2014-7151), the Dutch Pink Ribbon/A Sister's Hope (project number: Pink Ribbon-10074), Bayer AG Pharmaceuticals, Radiology (project number: BSP-DENSE), and Stichting Kankerpreventie Midden-West.

Correspondence to: Erik Verburg, MSc, Image Sciences Institute, University Medical Center Utrecht, Q.02.4.45, PO Box 85500, 3508 GA Utrecht, the Netherlands. E-mail: e.verburg-2@umcutrecht.nl; Kenneth G. A. Gilhuijs, PhD, Image Sciences Institute, Q.02.4.45, PO Box 85500, University Medical Center Utrecht, Utrecht University, Utrecht 3508 GA, the Netherlands. E-mail: k.g.a.gilhuijs@umcutrecht.nl.

Supplemental digital contents are available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.investigativeradiology.com).

ISSN: 0020-9996/20/5507–0438
DOI: 10.1097/RLI.0000000000000656

Women with extremely dense breasts (Breast Imaging Reporting and Data System [BI-RADS] class D), that is, breasts containing a large amount of fibroglandular tissue, have a 3 to 6 times higher risk of developing breast cancer than women with very fatty breasts. Moreover, these cancers are harder to detect on mammography due to the low contrast between fibroglandular tissue and tumor tissue and overlapping tissue.[1] Consequently, additional screening modalities, such as magnetic resonance imaging (MRI), have been proposed.

Magnetic resonance imaging is known to be a sensitive method to detect lesions. Several studies showed that additional MRI screening increases the number of detected malignancies.[2–5] However, MRI is also associated with lower specificity than mammography.[5,6] Moreover, MRI is more costly and time-consuming than mammography. The effectiveness of additional MRI for the screening of women with extremely dense breasts is the main research aim of the Dense Tissue and Early Breast Neoplasm Screening (DENSE) trial in the Netherlands. Within the framework of this randomized controlled trial, 4783 women with extremely dense breasts have been screened using additional MRI after a negative screening mammography.[7,8]

As anticipated, additional breast cancers were detected; in the first round of this trial, the cancer detection yield with MRI after negative mammography was 79 in 4783 women, or 16.5/1000 screens.[8] Subsequently, women in the MRI arm experienced a significantly lower number of interval cancers than those in the control arm.[8] However, in total, 454 women (9.5%) were referred for additional diagnostics after

MRI. BI-RADS 3 lesions led to recommendation for repeat MRI screening after 6 months and subsequent biopsy on indication. For women with BI-RADS 4 or BI-RADS 5 lesions, biopsy was indicated.

For women with BI-RADS 4 or BI-RADS 5 lesions, biopsy was indicated. As expected, the percentage of benign findings in BI-RADS 3– and BI-RADS 4–scored women was high. No malignant lesion was present in 97% of BI-RADS 3–scored women (146 of 150) and 79% of the BI-RADS 4–scored women (226 of 286). In BI-RADS 5–scored women, 17% (3 of 18) had no malignant lesions. Especially in BI-RADS 3 and BI-RADS 4–scored women, increased specificity would lead to reduced follow-up activities.

Reports on different, heterogeneous populations of women showed potential for computer-aided diagnosis (CAD) to improve the specificity of breast MRI.[9–11] To the best of our knowledge, no studies focused explicitly on a consecutively included screening population of asymptomatic women with extremely dense breasts and average risk.

Typically, CAD for breast MRI is based on dynamic contrast-enhanced T1-weighted images,[12] but combination with other sequences have been used as well (ie, multiparametric MRI).[11,13,14] In particular, high-temporal resolution dynamic contrast-enhanced series (fast-DCE),[13] diffusion-weighted imaging (DWI),[15] and T2-weighted imaging[16] have shown complementary value to discriminate between malignant and benign lesions.

To reduce the number of false-positive diagnoses in the MRI screening of women with extremely dense breasts, the aim of this study is to predict which BI-RADS 3 and BI-RADS 4 lesions are benign using multiparametric CAD.

MATERIALS AND METHODS

Study Population

Clinical data and MRIs were obtained during the first round of the DENSE trial. The DENSE trial has been described in detail elsewhere.[7] In short, this multicenter randomized controlled trial investigates the additional value of MRI screening in Dutch women with extremely dense breasts (ie, BI-RADS D and normal mammography). Written informed consent was obtained from all women before MRI screening. The trial was approved by the Dutch Minister of Health, Welfare, and Sport (2011/19 WBO, The Hague, the Netherlands). In this study, all image datasets were acquired between December 22, 2011, and January 22, 2016. All women with lesions that were scored as BI-RADS 3, 4, or 5 on MRI were included in the analysis described here. Some women with an indication for biopsy (31 of 331) did not undergo a biopsy because, for example, the lesion was not/no longer visible on additional imaging, the biopsy was technically not possible (in which case short-term follow-up imaging was applied), or the lesion was known to be benign from the patient records from another hospital.[8] The median age of the participants was 54 years (range, 49–75 years).

MRI Scan Acquisition

All breast MRI scans were acquired according to a fixed imaging protocol as described by Emaus et al.[7] In summary, the examinations were performed with a 3.0-T (Achieva or Ingenia) system from Philips or a 3.0-T (Trio, Verio, or Skyra) system from Siemens using a dedicated phased-array bilateral breast coil. Images were acquired in axial planes. The MRI protocol consisted of DWI, T1-weighted imaging without fat suppression, DCE-MR, and an optional T2-weighted sequence. Dynamic contrast-enhanced MR consisted of a high-spatial-resolution precontrast image, followed by a high-temporal-resolution series after contrast agent injection, followed by 4 or 5 high-spatial-resolution images. Fat suppression was optional during DCE-MR acquisition. The high-temporal-resolution series were acquired in 3.9- to 5.1-second intervals and consisted of 15 to 19 postcontrast acquisitions. Contrast agent was injected at a rate of 1 mL/s to a total dose of 0.1 mmol of macrocyclic gadolinium based contrast agent gadobutrol (Gadovist; Bayer AG, Leverkusen, Germany) per kilogram of body weight. Diffusion-weighted imaging was acquired with a minimum of 2 b-values and a maximal b-value of at least 800.[7]

Methods

A CAD model was developed and tested to predict whether lesions on MRI in women with extremely dense breasts are benign or malignant. The first stage of the CAD workflow was image processing (section 3.3.1) followed by automated calculation of features from all BI-RADS 3, 4, and 5 lesions (section 3.3.2). The features were used to train and validate the model using cross validation (section 3.3.3). These steps were repeated for a subset of images typically available in abbreviated MRI protocols,[9,17] that is, T2-weighted imaging, DWI and DCE-MRI consisting of high-temporal-resolution series, and one precontrast and one postcontrast image with a high spatial resolution.

Image Processing

Seven consecutive image processing steps were performed: (1) image registration of DCE-MR series, (2) lesion segmentation, (3) DCE-MRI scan normalization, (4) aorta segmentation, (5) chest wall segmentation and extraction of pectoral muscle intensity, (6) calculation of apparent diffusion coefficient (ADC), and (7) registration of lesion mask to ADC map.

1. Image registration of DCE-MRI scans: All postcontrast DCE-MRI scans with high spatial resolution were registered to their precontrast counterparts using a nonrigid B-spline transformation in a multiresolution scheme.[18]
2. Lesion segmentation: The semiautomated segmentation method proposed by Alderliesten et al[19] was used for lesion segmentation of mass lesions as well as nonmass lesions. Lesions were detected by breast radiologists associated with the DENSE trial, and whose experience ranged from 5 to 23 years.[8] A seed point was manually placed at or near the lesion by a technical physician (E.V.). Subsequently, constrained volume growing was performed in the DCE-MR series. This step resulted in a segmented lesion volume in 3D. Segmentations were reviewed by a trained breast radiologist (W.B.V.) and corrected when necessary by adding or replacing seed points.
3. DCE-MRI scan normalization: Although all images where acquired according to the screening protocol, some variations were present in the settings of the different MRI devices used, mainly flip angle and repetition time. Changes in intensity due to the inflow of contrast agent depend on these settings and may therefore differ between hospitals. Hence, we normalized intensities by calculating the signals that would be acquired at a standard flip angle and repetition time.[20,21] All DCE images with high spatial resolution were harmonized to a standard flip angle of 10 degrees and a standard repetition time of 3.78 milliseconds. All DCE images with high temporal resolution were harmonized to a standard flip angle of 10 degrees and a standard repetition time of 2.17 milliseconds.
4. Aorta segmentation: Contrast uptake speed in the lesion is related to contrast uptake in the descending aorta of the subject. Accordingly, the descending aorta was segmented in the DCE-MRI scans with high temporal resolution. The aorta was located on the basis of its tubular shape. We used the Hough transform to detect one circle with a diameter between 1 and 5 cm in each transversal slide in the last postcontrast series of the fast acquisition. A linear Hough transform was used to detect the main axis of the descending aorta. All found circles centered at the detected main axis were defined as the contours of the descending aorta.
5. Chest wall segmentation and extraction of pectoral muscle intensity: T2-weighted image intensities were normalized to the intensity of the pectoral muscle.[16] First, the pectoral muscle was automatically

detected in the T1-weighted images without fat suppression using dynamic programming.[22] Next, the detected chest wall was resampled to the dimensions of the T2-weighted image. The median intensities [MI(d)] of the voxels located at nearest distance d = 0, 1, 2 … n mm medial from the chest wall in the T2-weighted image were calculated. The pectoral muscle intensity was defined as MI(d) where the first local minimum or local maximum was present in MI(d).

6. Calculation of ADC: ADC values for all voxels in the DWI images of each subject were calculated using a linear least squares estimator based on QR decomposition. ADC values were computed using the nonweighted image (b-value = 0) in combination with all individual diffusion-weighted images (b-value > 0), which resulted in one ADC map per subject.
7. Registration of lesion mask to ADC map: DWI images are susceptible to artifacts such as geometric distortions due to magnetic susceptibility differences.[23] This geometric distortion was corrected by registration as follows: First, the T2-weighted images were registered to the corresponding DWI image with b-value 0 or 50. Nonrigid B-spline transformation in a multiresolution scheme was used.[18] Because the lesion mask is inherently aligned with the lesion on T2-weighted imaging, the transformation from T2 to DWI was applied to the lesion mask in order to align the mask with the lesion on ADC. Manual adjustment was applied by a technical physician (E.V.) when necessary.

Image registration was performed using Elastix (version: 4.7)[24]. MeVisLab (version 3.0; MeVis medical Solution AG, Bremen, Germany), in combination with Python (version 2.7; Python Software Foundation) with packages "numpy" (v1.15.1) and "scipy" (v1.1.0), was used for lesion segmentation, image normalization, and aorta segmentation. Chest wall segmentation and ADC map calculation were performed using MATLAB (v R2017a; Mathworks, Natick, MA).

Feature Extraction

In total, 49 features were calculated and used to train the CAD model. All 46 MRI-based features were obtained automatically. Twenty-two features describing morphology and contrast dynamics of the lesion were computed from the high-spatial DCE images.[12,25]

Six contrast uptake features were computed from the fast-DCE images using a method based on the work of Dalmiş et al.[13] Here, time-related features were expressed relative to the start of contrast uptake in the detected descending aorta. Nine ADC features and 9 T2 intensity features were computed (see Table, Supplemental Digital Content 1, http://links.lww.com/RLI/A516: Description of image features).

In addition to image features, 3 clinical features were considered in the model: BI-RADS score (3, 4, or 5), age, and body mass index (BMI).

Missing features (506 of 24,794) caused by missing images (n = 369), deviating imaging (n = 108), or missing clinical information (BMI only, n = 29) were multiply imputed (5 imputation sets).

A second feature set was extracted using only images that are available in abbreviated breast MRI protocols. Processing step 2, lesion segmentation, and feature extraction were repeated using only the first postcontrast images of the high-spatial-resolution dynamic image series. This feature set consisted of 36 image features and the same 3 clinical features.

Feature extraction was performed using MeVisLab (version 3.0; MeVis medical Solution AG, Bremen, Germany), Python (version 2.7; Python Software Foundation) with packages "numpy" (v1.15.1) and "scipy" (v1.1.0), and R (version 3.1.3, R Foundation for Statistical Computing, Vienna, Austria) with the packages "psych" (v1.5.8) and "Mice" (v2.25), which was used for data imputation.

Training and Validation

Outliers in feature values were defined as values deviating more than 3 standard deviations of the mean value. All feature values were normalized to values between 0 and 1. All BI-RADS 3, 4, and 5 lesions were used to train the model. The set was divided into 10 folds; each fold contained 7 or 8 malignant and 40 or 41 benign lesions to maintain the prevalence of malignancy observed in the DENSE study. To train the prediction model, 9 folds were used to fit a logistic regression model (the training set); the other fold was used as a validation set. BI-RADS 5 lesions were removed from the validation set, because in future application of the model, the gain of preventing false-positive diagnosis is expected to be low in this group. Cross-validation was repeated 10 times, each fold was used as validation set once. Before model fitting, feature values labeled as outliers in the training set were censored by clipping the extreme values.[26] To prevent overfitting, model weights were reduced for each fit using Ridge regression.[27] By iterating over all folds, cross-validated probabilities were obtained for all lesions. The results of 5 imputation sets were combined using Rubin rules.[28,29] The regularization parameter in the Ridge feature selection was determined using a second 10-fold cross-validation loop over the training data, using the deviance as performance measure. The regularization parameter was selected 1 standard error above the parameter with lowest error. Hence, we chose the simplest model whose accuracy was comparable with the best model.[30]

The posterior probabilities of the model to predict the presence of malignant disease in BI-RADS 3 and 4 lesions were summarized in an accuracy measure curve (AMC)[31] and receiver operator characteristic (ROC) curve. An AMC shows the percentage of correctly predicted malignant lesions and correctly predicted benign lesions for a range of values of the probability threshold (pt) using:

$$Sensitiviy(pt) = \frac{TP(pt)}{TP(pt) + FN(pt)} \quad \text{Equation 1}$$

$$Specificity(pt) = \frac{TN(pt)}{TN(pt) + FP(pt)} \quad \text{Equation 2}$$

$$PPV\ (pt) = \frac{TP(pt)}{TP(pt) + FP(pt)} \quad \text{Equation 3}$$

$$NPV(\mathrm{pt}) = \frac{TN(pt)}{TN(pt) + FN(pt)} \quad \text{Equation 4}$$

where TP is the number of true-positives, TN the number of true-negatives, FP the number of false-positives, and FN the number of false-negatives for each probability threshold. In each decision curve, 3 operating points were selected, at a sensitivity level of 100%, 99%, and 95%, respectively. The models were compared using the McNemar χ^2 test; a P value less than 0.05 was considered significant.

Training and validation were performed using R (version 3.1.3, R Foundation for Statistical Computing, Vienna, Austria) with the packages "psych" (v1.5.8), "glmnet" (v2.0-5), and "pROC" (v1.8).

RESULTS

In the total screening population of 4783 women, 81 malignant lesions were found in 79 women and 444 benign lesions were found in 390 women. Fifteen women had both a malignant and a benign lesion. Four malignant lesions in 3 women were excluded because the images were not available. Fifteen benign lesions were excluded: images of 9 lesions were not available for this study, one lesion was imaged using a deviant MRI protocol, and 5 lesions could not be examined due to imaging artifacts affecting correct segmentation or feature extraction. Image artifacts were caused by movement of the woman during imaging. The deviant images or artifacts did not alter the ability of the radiologist to score the images; however, the images were unusable

TABLE 1. Lesion Types of BI-RADS 3, 4, and 5 Lesions

Lesion Types	429
Benign Lesions	
Adenomyoepithelioma	2
Adenosis	24
Apocrine metaplasia	14
Atypical ductal hyperplasia	5
Cholesterol crystal	1
Cylindrical cell metaplasia	1
Cyst	8
Fat necrosis	1
Fibroadenoma	40
Fibrosis	35
Hemangioma	2
Lobular carcinoma in situ*	4
Lipoma	3
Lobular hyperplasia	1
Lobular neoplasia	4
Lobulitis	3
Lymph node	5
Mastopathy	27
Normal breast tissue	23
Papilloma	18
Usual ductal hyperplasia	32
Unknown†	176
Malignant lesions	**77**
Ductal carcinoma in situ	13
Invasive ductal carcinoma	35
Invasive ductal lobular carcinoma	5
Invasive intracystic papillary carcinoma	2
Invasive lobular carcinoma	13
Invasive mucinous carcinoma	1
Invasive tubular carcinoma	8

Results were obtained from the pathology reports after biopsy.

*In the DENSE trial, LCIS is considered a benign lesion,[8] conforming to the Dutch guidelines.[32]

†No biopsy result was available for these lesions. No biopsy performed after BI-RADS 3 score (n = 153); the lesion was not/no longer visible on additional imaging (n = 16); biopsy result was unknown (n = 7).

for the automated method presented. In summary, 429 benign and 77 malignant lesions were available for this study (Tables 1–3). Most lesions (293) were scored BI-RADS 4, whereas 194 lesions were scored BI-RADS 3, and 19 lesions were scored BI-RADS 5 (Table 2).

The AMC and corresponding ROC curve (Fig. 1) show feasibility to increase the specificity using CAD. An AMC and ROC curve were obtained for the subgroup of all BI-RADS 3 and BI-RADS 4 lesions. The presented model outputs a probability of malignancy for each lesion. Cutoff thresholds in this probability define the sensitivity and specificity of the model. We chose 3 cutoff thresholds corresponding to sensitivity 100%, 99%, and 95%. The cutoff thresholds are shown in the AMCs (Fig. 1). Corresponding specificity at each threshold is shown in Table 4.

The full multiparametric model classified 176/425 (41.5%) of the false-positive BI-RADS 3– and BI-RADS 4–scored lesions as benign without missing a malignancy. Of the total group of lesions referred to additional MRI or biopsy, 425/487 (87.3%) were false-positive. With additional CAD used before referral, this fraction may be reduced to 311/487

TABLE 2. Overview of All Lesions Used for Development of the CAD Model Stratified by BI-RADS Score and Mass- or Non–Mass-Enhancing Lesions

BI-RADS Score	No. Benign Lesions	No. Malignant Lesions
3		
Mass	105	3
NME	84	2
4		
Mass	168	47
NME	68	10
5*		
Mass	4	14
NME	0	1
Total	**429**	**77**

*Only used for model training, not for model testing.

NME, non–mass-enhancing lesions.

(63.9%). For the abbreviated protocol model, the referrals would be 376 instead of 487 (77.2%). Examples of lesions that were false-positive and correctly identified as such by CAD are shown in Figure 2 and Figure 3.

DISCUSSION

From a screening population of 4783 women, MRI scans of 77 malignant breast lesions and 429 benign lesions were used to create a multiparametric CAD model based on Ridge regression, to identify benign disease with high certainty. The model may have potential to reduce follow-up on benign BI-RADS 3 and BI-RADS 4 lesions: 41.5% reduction for the full multiparametric protocol and 26.2% for the abbreviated protocol model, without missing a malignant lesion.

Although the performance of the full-protocol model and the abbreviated-protocol model is comparable in terms of AUC (0.85 vs 0.84), the number of detected benign lesions without missing malignant lesions was observed to be higher in the full-protocol model ($P < 0.01$). These results suggest that the high-resolution postcontrast images contain information to increase the specificity to identify benign disease at high sensitivity. We observed comparable performance between mass and nonmass lesions, indicating that the features accurately describe both types of lesions (Supplemental Digital Content 2, http://links.lww.com/RLI/A517: Performance in mass- and non–mass-enhancing lesions).

To our knowledge, this study is the first to apply a multiparametric CAD model in unselected homogeneous data obtained from a multicenter screening trial in women with extremely dense breasts. The performance of presented models (AUC of 0.85 ± 0.04 and 0.84 ± 0.04) based on DCE, fast-DCE, ADC, T2, and clinical data is on par to that of other published methods in other study populations. Dalmiş et al[9] used deep learning on fast-DCE, T2, and DWI, and obtained an AUC of 0.852. Other authors designed multiparametric models

TABLE 3. Comparison Between Malignant and Benign Lesions Used for This Study

	Benign	Malignant	*P*
No. lesions	429	77	
Lesion volume, cm^3	0.18 (0.10–0.36)	0.33 (0.16–0.77)	<0.001
BMI	22.25 (20.75–24.01)	22.86 (21.48–24.95)	0.013
Age, y	53.05 (50.90–56.90)	54.80 (51.30–61.70)	0.008

For statistical comparison, the Kruskal-Wallis test was used; $P < 0.05$ was considered significant. Median feature values and interquartile range are shown.

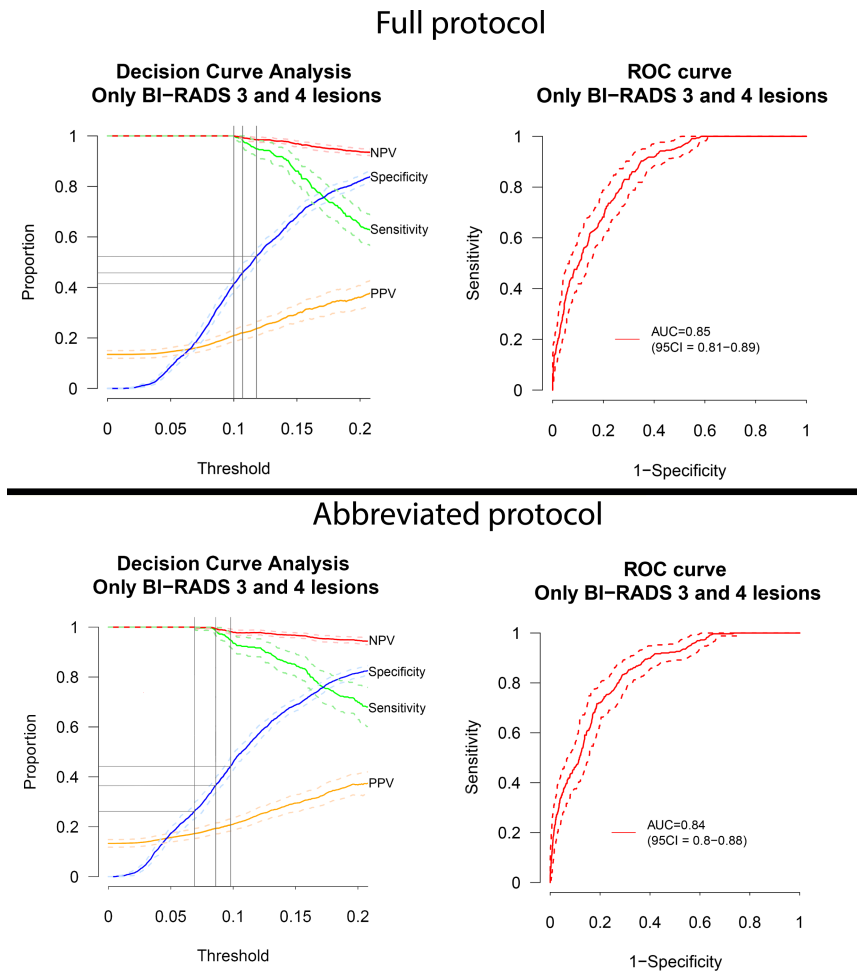

FIGURE 1. Accuracy measure curve of the CAD model and corresponding ROC curve using all MRI series (top row) and abbreviated imaging series (bottom row). The blue curve denotes specificity; the yellow curve, the positive predictive value (PPV); the green curve, the sensitivity; and the red curve, the negative predictive value (NPV). One standard deviation corrected for multiple imputation using Rubin rules is shown using dashed curves. The vertical gray lines indicate the cutoff threshold probabilities (pt) corresponding to sensitivity 100%, 99.0%, and 95.0% from left to right.

using fewer image sequences for feature extraction, for example, DCE and T2-weighted images,[11,33,34] resulting in AUCs of 0.88 ± 0.01,[11] 0.83 ± 0.03,[33] or 0.85 ± 0.03.[34] Using only DCE yielded comparable results (AUC of 0.85).[35] However, the above results might not be directly comparable to our results because they were based on single institution data. In addition, the current study omitted BI-RADS 2 and 5 lesions. The rationale for this omission is that the problem of false-positives does not occur in BI-RADS 2 and BI-RADS 5. BI-RADS 2 lesions are not referred, and by definition, BI-RADS 2 lesions are nearly always benign. BI-RADS 5 lesions are nearly always malignant. By omitting

TABLE 4. Overview of Correctly Classified Benign BI-RADS 3 and 4 Lesions and Corresponding Levels of Correctly Classified Malignant Lesions for Both Models

Correctly Classified Malignant Lesions	Correctly Classified Benign BI-RADS 3 and 4 Lesions		
	Full Protocol	Abbreviated Protocol	*P*
100.0%	41.5% ± 3.2%	26.2% ± 3.2%	<0.01
99.0%	45.8% ± 3.5%	36.6% ± 3.0%	<0.01
95.0%	52.4% ± 3.1%	44.3% ± 3.6%	<0.01

Results denote mean ± 1 standard deviation. Models were compared using the McNemar χ^2 statistic.

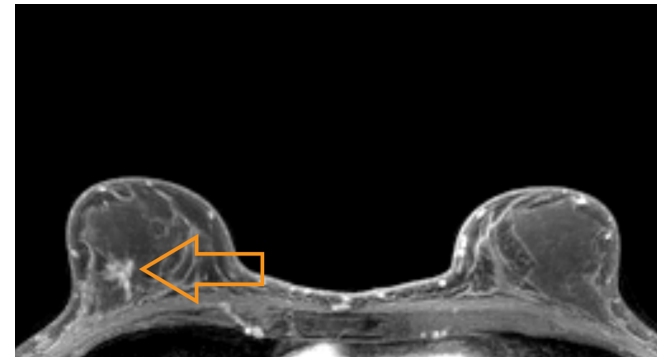

FIGURE 2. Maximum intensity projection of a 14-mm false-positive lesion in a 59-year-old woman who was referred to biopsy. The BI-RADS 4 classified lesion (right breast) was a benign fibrotic lesion (arrow). The computer-aided diagnosis correctly classified it as benign with probability of malignancy of 2.5%.

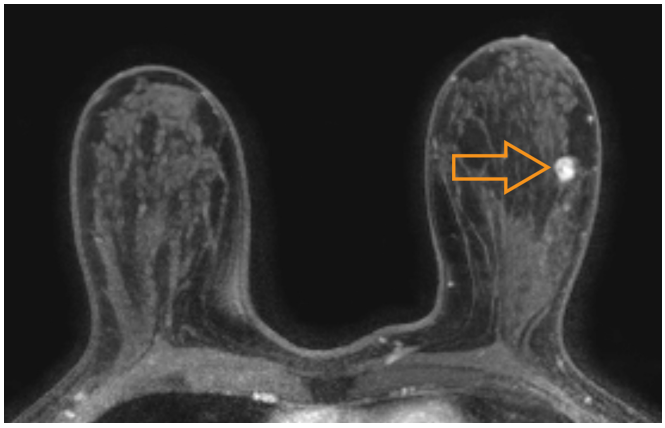

FIGURE 3. Maximum intensity projection of a 10-mm false-positive lesion in a 52-year-old woman who was referred to biopsy. The BI-RADS 4 classified lesion (left breast) was a benign fibroadenoma (arrow). The computer-aided diagnosis correctly classified it as benign with probability of malignancy of 2.1%.

these categories, the CAD is tested on the most difficult and clinically relevant cases.

In this study, the risk of overfitting was reduced using Ridge regression.[36] Features with the largest regression weights were signal-enhancing ratio, top washout, volume uptake, and volume washout from the DCE image series; the maximum slope and general slope from the fast-DCE; and the 75th percentile of ADC values in the lesion. T2 features did not have high weighting in the model.

We did not use deep learning because the number of malignant lesions was relatively small for such an approach. Deep learning can outperform linear regression methods when the number of training data is large enough to avoid overfitting.[11] Currently, however, the literature does not indicate a clear benefit of deep learning over radiomics for this problem, other than that deep-learning models are less time-consuming to construct. The largest study on deep learning to discriminate between benign and malignant disease on MRI uses 1294 cases,[11] and yields comparable performance (AUC of 0.88). A potential risk of deep learning is, however, that the millions of parameters that describe the data may cause unnoticed bias in the detection of malignant disease in populations for which the model was not explicitly trained.

Although all MRI data were acquired according to the same protocol, variation was introduced between institutions because MRI scanners from different vendors were used. Moreover, some MRI settings varied across hospitals, for example, the use of fat suppression, flip angle, and repetition time. We used a data harmonization step between MRI scanners to counter the effect of some of these variations.

This study also has some limitations. We were not able to validate the method in an unseen dataset, but used cross-validation. In future research, the CAD model should be validated in an independent population of women with extremely dense breasts. Another potential limitation is that we used lesions found only during the first round of the screening trial. We have not yet investigated whether the machine-extracted phenotype of lesions detected in subsequent, or incident, screening rounds is representative of that detected in the first round. For instance, lesions may be smaller on average in subsequent screening rounds, and perhaps also more aggressive. The tumors in the first, or prevalent round, may comprise relatively slow-growing, less aggressive tumors that have been present for a long time.

We describe computerized analysis of MRI scans with BI-RADS score of the radiologist as input. Computer-aided diagnosis is, however, typically implemented as an aid to radiologists, using the computer as second opinion. This interaction has not yet been investigated.

In conclusion, we developed a CAD method based on Ridge regression to identify benign lesions with high certainty in multiparametric breast MR screening of extremely dense breasts, thus pursuing to reduce the number of recalls on benign lesions. Using internal validation, the method showed potential to reduce referral of benign BI-RADS 3 and BI-RADS 4 lesions without loss of sensitivity.

ACKNOWLEDGMENTS

The authors acknowledge the study participants for their contributions: H.M. Chan, M.J. van Rijssel, Dr B.H.M. van der Velden, and Dr A. de Luca for assistance in the MR physics aspects of this study; B.M. den Dekker and S.V. de Lange for assistance in data management; and Dr P.J. van Diest for assistance in the pathological aspects of this study.

REFERENCES

1. Carney PA, Miglioretti DL, Yankaskas BC, et al. INdividual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Ann Intern Med*. 2003;138:168–175.
2. Berg WA, Zhang Z, Lehrer D, et al. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *JAMA*. 2012;307:1394–1404.
3. Kuhl CK, Schrading S, Strobel K, et al. Abbreviated breast magnetic resonance imaging (MRI): first Postcontrast subtracted images and maximum-intensity projection—a novel approach to breast cancer screening with MRI. *J Clin Oncol*. 2014;32:2304–2310.
4. Kuhl CK, Strobel K, Bieling H, et al. Supplemental breast MR imaging screening of women with average risk of breast cancer. *Radiology*. 2017;283:361–370.
5. Saadatmand S, Geuzinge HA, Rutgers EJ, et al, FaMRIsc study group. MRI versus mammography for breast cancer screening in women with familial risk (FaMRIsc): a multicentre, randomised, controlled trial. *Lancet Oncol* 2019;20:1136–1147.
6. Menezes GL, Knuttel FM, Stehouwer BL, et al. Magnetic resonance imaging in breast cancer: a literature review and future perspectives. *World J Clin Oncol*. 2014;5:61–70.
7. Emaus MJ, Bakker MF, Peeters PH, et al. MR imaging as an additional screening modality for the detection of breast cancer in women aged 50-75 years with extremely dense breasts: the DENSE trial study design. *Radiology*. 2015;277:527–537.
8. Bakker MF, de Lange SV, Pijnappel RM, et al. Supplemental MRI screening for women with extremely dense breast tissue. *N Engl J Med*. 2019;381:2091–2102.
9. Dalmis MU, Gubern-Mérida A, Vreemann S, et al. Artificial intelligence-based classification of breast lesions imaged with a multiparametric breast MRI protocol with ultrafast DCE-MRI, T2, and DWI. *Invest Radiol*. 2019;54:325–332.
10. Zhang M, Horvat JV, Bernard-Davila B, et al. Multiparametric MRI model with dynamic contrast-enhanced and diffusion-weighted imaging enables breast cancer diagnosis with high accuracy. *J Magn Reson Imaging*. 2019;49:864–874.
11. Truhn D, Schrading S, Haarburger C, et al. Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI. *Radiology*. 2019;290:290–297.
12. Gilhuijs KG, Deurloo EE, Muller SH, et al. Breast MR imaging in women at increased lifetime risk of breast cancer: clinical system for computerized assessment of breast lesions—initial results. *Radiology*. 2002;225:907–916.
13. Dalmış MU, Gubern-Mérida A, Vreemann S, et al. A computer-aided diagnosis system for breast DCE-MRI at high spatiotemporal resolution. *Med Phys*. 2016;43:84–94.
14. Tahmassebi A, Wengert GJ, Helbich TH, et al. Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. *Invest Radiol*. 2019;54:110–117.
15. Kuroki Y, Nasu K, Kuroki S, et al. Diffusion-weighted imaging of breast cancer with the sensitivity encoding technique: analysis of the apparent diffusion coefficient value. *Magn Reson Med Sci*. 2004;3:79–85.
16. Ballesio L, Savelli S, Angeletti M, et al. Breast MRI: are T2 IR sequences useful in the evaluation of breast lesions? *Eur J Radiol*. 2009;71:96–101.
17. Chhor CM, Mercado CL. Abbreviated MRI protocols: wave of the future for breast cancer screening. *AJR Am J Roentgenol*. 2016;208:284–289.
18. Gubern-Mérida A, Martí R, Melendez J, et al. Automated localization of breast cancer in DCE-MRI. *Med Image Anal*. 2015;20:265–274.
19. Alderliesten T, Schlief A, Peterse J, et al. Validation of semiautomatic measurement of the extent of breast tumors using contrast-enhanced magnetic resonance imaging. *Invest Radiol*. 2007;42:42–49.

20. Haacke EM, Filleti CL, Gattu R, et al. New algorithm for quantifying vascular changes in dynamic contrast-enhanced MRI independent of absolute T1 values. *Magn Reson Med*. 2007;58:463–472.
21. van Rijssel MJ, Pluim JPW, Chan HM, et al. Correcting time-intensity curves in dynamic contrast-enhanced breast MRI for inhomogeneous excitation fields at 7T. *Magn Reson Med*. 2019; In press.
22. Verburg E, Wolterink JM, de Waard SN, et al. Knowledge-based and deep learning-based automated chest wall segmentation in magnetic resonance images of extremely dense breasts. *Med Phys*. 2019;46:4405–4416.
23. Chilla GS, Tan CH, Xu C, et al. Diffusion weighted magnetic resonance imaging and its recent trend-a survey. *Quant Imaging Med Surg*. 2015;5: 407–422.
24. Klein S, Staring M, Murphy K, et al. Elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging*. 2010;29:196–205.
25. Gilhuijs KG, Giger ML, Bick U. Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging. *Med Phys*. 1998;25: 1647–1654.
26. Hastings C, Mosteller F, Tukey JW, et al. Low moments for small samples: a comparative study of order statistics. *Ann Math Statist*. 1947;18:413–426.
27. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Dent Tech*. 1970;12:55–67.
28. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: Wiley; 2004.
29. Marshall A, Altman DG, Holder RL, et al. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol*. 2009;9:57.
30. Breiman L. *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall/CRC; 1984.
31. Vos S, Elias SG, van der Groep P, et al. Comprehensive proteomic profiling-derived immunohistochemistry-based prediction models for *BRCA1* and *BRCA2* Germline mutation-related breast carcinomas. *Am J Surg Pathol*. 2018;42:1262–1272.
32. Smorenburg CH. Borstkanker Landelijke richtlijn. Version 1.0. Integraal kankercentrum Nederland (IKNL). Available at: https://www.oncoline.nl/index.php?pagina=/richtlijn/item/pagina.php&id=41189&richtlijn_id=1069. Accessed September 1, 2019.
33. Gallego-Ortiz C, Martel AL. Using quantitative features extracted from T2-weighted MRI to improve breast MRI computer-aided diagnosis (CAD). *PLoS One*. 2017;12:e0187501.
34. Bhooshan N, Giger M, Lan L, et al. Combined use of T2-weighted MRI and T1-weighted dynamic contrast-enhanced MRI in the automated analysis of breast lesions. *Magn Reson Med*. 2011;66:555–564.
35. Deurloo EE, Muller SH, Peterse JL, et al. Clinically and mammographically occult breast lesions on MR images: potential effect of computerized assessment on clinical Reading. *Radiology*. 2005;234:693–701.
36. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res* 2019; 28:2455–2474.