Prediction of Malignancy by a Radiomic Signature From Contrast Agent-Free Diffusion MRI in Suspicious Breast Lesions Found on Screening Mammography.

Sebastian Bickelhaupt, MD,1 Daniel Paech, MD,1 Philipp Kickingereder, MD,1,2 Franziska Steudle,1 Wolfgang Lederer, MD,3 Heidi Daniel, MD,4 Michael Götz, PhD,5 Nils Gähler,5 Diana Tichy, PhD,6 Manuel Wiesenfarth, PhD,6 Frederik B. Laun, PhD,7 Klaus H. Maier-Hein, PhD,5 Heinz-Peter Schlemmer, MD, PhD,1 and David Bonekamp, MD1*

Purpose: To assess radiomics as a tool to determine how well lesions found suspicious on breast cancer screening X-ray mammography can be categorized into malignant and benign with unenhanced magnetic resonance (MR) mammography with diffusion-weighted imaging and T2-weighted sequences.

Materials and Methods: From an asymptomatic screening cohort, 50 women with mammographically suspicious findings were examined with contrast-enhanced breast MRI (ceMRI) at 1.5T. Out of this protocol an unenhanced, abbreviated diffusion-weighted imaging protocol (ueMRI) including T2-weighted, (T2w), diffusion-weighted imaging (DWI), and DWI with background suppression (DWIBS) sequences and corresponding apparent diffusion coefficient (ADC) maps were extracted. From ueMRI-derived radiomic features, three Lasso-supervised machine-learning classifiers were constructed and compared with the clinical performance of a highly experienced radiologist: 1) univariate mean ADC model, 2) unconstrained radiomic model, 3) constrained radiomic model with mandatory inclusion of mean ADC.

Results: The unconstrained and constrained radiomic classifiers consisted of 11 parameters each and achieved differentiation of malignant from benign lesions with a .632 + bootstrap receiver operating characteristics (ROC) area under the curve (AUC) of 84.2%/85.1%, compared to 77.4% for mean ADC and 95.9%/95.9% for the experienced radiologist using ceMRI/ueMRI.

Conclusion: In this pilot study we identified two ueMRI radiomics classifiers that performed well in the differentiation of malignant from benign lesions and achieved higher performance than the mean ADC parameter alone. Classification was lower than the almost perfect performance of a highly experienced breast radiologist. The potential of radiomics to provide a training-independent diagnostic decision tool is indicated. A performance reaching the human expert would be highly desirable and based on our results is considered possible when the concept is extended in larger cohorts with further development and validation of the technique.

Level of Evidence: 1
Technical Efficacy: Stage 2

J. MAGN. RESON. IMAGING 2017;46:604–616

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.25606
Received Oct 12, 2016, Accepted for publication Dec 7, 2016.

*Address reprint requests to: D.B., Department of Radiology (E010), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. E-mail: d.bonekamp@dkfz-heidelberg.de

From the 1Department of Radiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; 2Department of Neuroradiology, University of Heidelberg Medical Center, Heidelberg, Germany; 3Radiological Clinic at the ATOS Clinic Heidelberg, Heidelberg, Germany; 4Radiology Center Mannheim (RZM), Mannheim, Germany; 5Junior Group Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany; 6Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany; and 7Department of Medical Physics in Radiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

Additional Supporting Information may be found in the online version of this article.
Breast cancer screening programs have been repeatedly criticized for the harms related to false-positive findings on screening X-ray mammography. While many false-positive findings are corrected during the screening clarification process that is initiated in cases of suspicious lesions detected on radiographic mammography by means of noninvasive diagnostic measures such as ultrasound, clinical examination, and magnification mammography in order to more precisely characterize the lesion, still about 50% of all biopsies indicated after this noninvasive clarification will eventually reveal a benign finding. For women who participate in breast cancer screening programs undergoing 10 screening rounds (eg, in Germany the breast cancer screening is offered every 2 years, thus 10 screening rounds cover a time span of 20 years), a risk of 1.8–6.3% to undergo biopsy related to a false-positive finding has been reported. Noninvasive diagnostic measures complementing radiographic mammography are candidates to reduce the number of false-positive findings triggering invasive procedures in breast cancer screening participants. In terms of screening feasibility, ultrasound elastography and unenhanced abbreviated breast magnetic resonance imaging (MRI) currently appear to have potential for added benefit during noninvasive work-up. Recently, contrast-free breast MRI protocols with high-quality diffusion-weighted imaging (DWI) have shown promise in the diagnostic decision-making of breast lesions. The combination of DWI with morphological \(T_2\)-weighted imaging sequences has been reported to provide reliable characterization of breast lesions. DWI is increasingly being incorporated into breast MRI, as it has shown promise in improved differentiation of benign and malignant lesions and noncontrast detection of breast cancer.

A further measure to improve this methodic approach may lie in its combination with a radiomic approach. Radiomics is a recently emerging field in radiology that uses advanced image processing techniques to extract a large number of descriptive parameters from imaging data that can be used to build predictive models of clinical outcome parameters. It is hypothesized that by extracting a large number of quantitative imaging parameters from imaging data, information hidden from confident assessment by the human eye can be made accessible and, due to the large number of parameters, can provide a noninvasively characterizable imaging phenotype that complements or partially predicts the disease genotype. The radiomic parameters currently in use are based on mathematical definitions that describe descriptive statistics, shape, and texture information about the tissue under investigation. Prior studies have reported promising performance of texture features and radiomics with and without the use of machine learning in the prediction of tumor response from MR mammography (MRM). Most MRM radiomics studies so far have focused specifically on contrast-enhanced MR sequences, and have evaluated tumor response to neoadjuvant chemotherapy or prediction of histological subtype, OncotypeDX risk categories, early metastasis, molecular subtypes, gene and protein expression, and commonly used gene assays.

The purpose of this study was to assess the ability of radiomics to determine how well lesions found suspicious on breast cancer screening X-ray mammography can be categorized into malignant and benign with unenhanced MR mammography with DWI and \(T_2\)-weighted sequences, and compare this performance to that of an experienced radiologist interpretation and to the established diffusion parameter apparent diffusion coefficient (ADC).

**Materials and Methods**

Prospective data collection was approved by local and governmental ethics committees (ethics approval numbers S-151/2014 and B-FF-2014-069) and informed consent was obtained from all patients prior to entering this study. This retrospective subgroup analysis included 50 consecutive patients (57.6 years \(SD \pm 6.4\) years) out of the study cohort. All patients in the study cohort presented with a suspicious X-ray screening mammogram (category 4 or 5 classification in the Breast Imaging Reporting and Data System [BI-RADS]) and indication for breast biopsy. Reading of screening X-ray mammography was performed double-blinded by screening certified radiologists. Prior to MRI, all women underwent the regular screening work-up process including ultrasound, clinical examination, and, if necessary, a magnification mammography for further characterization of the lesions. The examined cohort was subject to a prior study focusing on diagnostic performance of unenhanced MRM in the hands of expert radiologists and the added benefit of maximum intensity projections (MIPs) and image fusion techniques; however, radiomic analyses in this group have not been previously performed. Baseline epidemiological and clinical characteristics including tumor location, pathological findings, and BI-RADS assessment are shown in Supplementary Table 1. Half (25 out of 50) of the patients were found to have a malignant lesion (50%) on histopathology. In one patient the biopsy yielded inconclusive results despite sufficient cellular material and full histopathological analysis including immunohistochemistry; clinically the risk of malignancy was considered low enough to support the patient decision of foregoing repeat biopsy and continuing imaging follow-up. This case was considered a benign diagnosis for this analysis.

**MRI**

MR images were acquired using a 1.5T scanner (Philips Ingenia, Best, The Netherlands) using a 2-channel breast coil with additive elements in the MRI table. All women were scanned in prone position with the breast not compressed but softly supported using foam material, as previously reported.

**Sequences**

The contrast-enhanced full-breast MRI protocol (cEMRI) was acquired as a full diagnostic breast MRI protocol including...
noncontrast-enhanced \( T_1 \)-weighted and \( T_2 \)-weighted, DWI with background suppression (DWIBS) and contrast-enhanced (gadobenate dimeglumine, 0.1 mmol/kg, Multihance, Bracco, Milan, Italy) sequences with subtraction images, as previously reported.\(^{11}\) The following sequences were extracted for the abbreviated unenhanced breast MRI protocol (ueMRI): \( T_2 \)-weighted turbo-spin echo (T2TSE), DWIBS, and DWI (details and sequence parameters are given in Supplementary Table 2). A high b-value of 1500 s/mm\(^2\) was acquired for diffusion sequences in order to increase lesion conspicuity compared to regular breast parenchyma. ADC maps were calculated from the DWIBS (ADC\(_{\text{DWIBS}}\) from DWIBS b = 0 s/mm\(^2\) and b = 1500 s/mm\(^2\) images) and DWI sequences (ADC\(_{\text{DWI}}\) from DWI b = 0 s/mm\(^2\) and b = 1500 s/mm\(^2\) images) using a monoexponential fit.\(^{26}\)

**Image Postprocessing and Analysis**

Clinical MR BI-RADS assessment was performed prospectively by a specialized breast imaging radiologist (W.L., >20 years of experience in breast imaging) in two reading rounds. First, the expert radiologist assessed an abbreviated protocol excluding contrast-enhanced sequences and utilizing only \( T_2 \)-weighted images, DWIBS, DWI sequences and ADC maps calculated from DWIBS and DWI images in a similar manner as previously reported.\(^{11}\) In short, initially the DWI/DWIBS sequences were used to determine the presence of a lesion (appearance of restricted diffusion over background signal). Since lesion morphology cannot be assessed fully on DWI sequences, the \( T_2 \)-weighted images were then used to further assess lesion morphology according to the ACR MR BI-RADS scheme. Afterwards, the expert radiologist was asked to score the lesions detected according to a Likert-like score as previously described with a score of 1 indicating no suspicious lesion and 5 indicating a highly suspicious lesion.\(^{11}\)

After conducting the evaluation of the abbreviated ueMRI protocol, the expert reader was given access to the full diagnostic protocol (including contrast-enhanced sequences and subtraction images) that was then interpreted directly following the interpretation of the abbreviated protocol, according to the recommendations by the ACR MR BI-RADS system. An MR BI-RADS (ceMRI) or Likert (ueMRI) score of 4 or 5 was interpreted as a prediction of malignancy.

T2TSE, DWIBS1500, ADC\(_{\text{DWIBS}}\), and ADC\(_{\text{DWI}}\) maps were upsampled to the \( T_1 \)-DCE resolution (0.7 mm in-plane resolution and 1 mm slice thickness) using the FMRIB software library (FSL, fsl.fmrib.ox.ac.uk/fsl) using trilinear interpolation, to allow the radiomic feature extraction to operate at the same scale for all imaging sequences and to model the interpolation provided by the picture archiving and communication system (PACS) used by the radiologist.\(^{25}\) \( T_1 \)-DCE images were used only for this step, to generate high-resolution and spatially matched datasets from the other sequences. All further processing did not include information from the contrast-enhanced sequences. Steps of image postprocessing and analysis are summarized and illustrated in Fig. 1. 3D segmentations of the MR index lesions were generated manually (S.B., D.P., F.S.) using the medical imaging toolkit software (MITK, www.mitk.org)\(^{26}\) and performed separately on T2TSE images and DWIBS1500 images.\(^{27}\) Lesions were matched to the lesions reported by the expert readers by review of the mammograms and correlation to written documentation in the study score sheets and screenshot images performed during initial reading of the cases to assure the outlined lesions were identical to those used for BI-RADS reporting. One operator (F.S.) performed segmentations of the background parenchyma on DWIBS1500 images and normal-appearing fat on the T2TSE image, which were used to normalize the MR intensities of the corresponding images in terms of lesion to background ratio. ADC-maps, being already in quantitative units, were not normalized.

Radiomic feature calculations were performed with the medical imaging toolkit (MITK, www.mitk.org).\(^{26}\) Within each VOI, 1) 17 first-order features (FO), 2) nine volume and shape features (VSF), and 3) 162 texture features (TF) were calculated. VSF depends on the binary information of the segmentation mask only, while FO and TF reflect the intensities of the examined imaging sequence. Intensity normalized DWIBS1500 and T2 images, and ADC\(_{\text{DWIBS/DWI}}\) images yielded 716 FO and TF. Together with nine VSF, this resulted in a total of 725 radiomics features for each patient. To limit the initial parameter space, DWI1500 images were not entered into the feature set as they provide non-quantitative information that was considered redundant in comparison with DWIBS1500. FO represent the voxel intensities by first-order statistics, including mean, standard deviation, kurtosis, skewness, uniformity, energy, and entropy. VSF characterize the VOI shape by metrics such as compactness, maximal 3D diameter, spherical disproportion, surface area, volume, and surface to volume ratio. TF were based on both co-occurrence and run-length-based features. Co-occurrence features were calculated on the basis of gray level co-occurrence matrices (GLCM) and included Haralick features, while run-length-based features represent the structure of an image region characterized by a gray level run length matrix (GLRLM). A detailed definition of the included radiomics parameters is given in the Data Supplement S1.

**BIOPSY.** All participants of the study received a biopsy (core needle breast biopsy) of the lesion found on screening X-ray mammography using either ultrasound or stereotactic radiographic guidance. Histopathological results served as the standard of reference regarding diagnostic accuracy.

**Statistical Analysis**

We applied a Lasso regularized logistic regression model. Thereby, the negative log-likelihood with an L1-penalty is minimized

\[
\text{minimize}_{\beta_0, \beta} \left\{ -\frac{1}{N} \sum \beta_0, \beta(y, X) + \lambda \| \beta \|_1 \right\}
\]

where \( \ell \) is the log likelihood, \( y \) and \( X \) are the response and design matrix, respectively, \( \beta_0 \) is the intercept, and \( \beta \) is the coefficient vector of all features. The second term is the L1-penalty controlled by tuning parameter \( \lambda \). Thereby, \( \lambda \) controls the amount of shrinkage of the regression coefficients towards zero. Making \( \lambda \) sufficiently large will cause some of the coefficients to be exactly zero and thus we obtain a sparse solution with implicit feature selection. \( \lambda \) is selected during the fitting process using cross-validation to optimize prediction performance. The choice of Lasso was made due to its computational efficiency and well-studied properties. It is well known that a regularized regression model provides better...
prediction with respect to mean squared error than ordinary regression. Further, regularization allows us to include more features than observations. Note, however, that the Lasso selects at most $n$ features. Model selection consistency of the Lasso has been shown by Zhao and Yu.28 Three models were built: 1) a univariate model based on mean ADCDWIBS (without regularization), 2) a model with unconstrained parameter selection, and 3) a model into which mean ADCDWIBS was included as a mandatory variable to permit assessment which additional variables extend the performance of ADC, and to determine if inclusion of such prior knowledge can lead to improved parameter selection.

The prediction performance of the classifier was validated by .632+ bootstrap techniques.29 Within this approach the data were repeatedly split into training and test sets and within each split the tuning parameter was chosen by cross-validation using the training data in each given split. Instead of a single data split into a training and a validation set, many of such splits are performed in bootstrap resampling to deal with several issues arising in single data splits such as loss of power and selecting “fortunate” or “unfortunate” splits by chance. Classification performance was then estimated by averaging over the prediction performances in the out-of-bag sets. The .632+ bootstrap estimate is a weighted average between the apparent error and the leave-one-out bootstrap estimate of the prediction error. In all, 1000 bootstrap samples were used. Such a nested resampling strategy for validation of the classification performance is considered crucial to reduce overfitting and thus overly optimistic performance estimates.

Quantiles of the leave-one-out bootstrap distributions are used to assess the uncertainty in the AUC estimate. Note that they are centered around the more conservative leave-one-out bootstrap estimate of the AUC.

All statistical analyses were carried out using R v. 3.3.0 (R Foundation for Statistical Computing, Vienna, Austria).30 Lasso regularized logistic regression model was estimated using R package glmnet 2.0-5, model validation was carried out using packages ModelGood 1.0.9 and peperr 1.1-7. A $P$ of 0.05 was considered statistically significant.

**RESULTS**

Mean lesion size on mammography was 11 mm (range 4–60 mm). No significant differences were found for lesion size (median 11 mm for both groups, $P = 0.129$) between the patient groups with malignant and benign lesions; however, patients with benign lesions were statistically younger ($P = 0.015$, median 53 years, 25%/75% percentile 51–59 years) compared to patients with malignant lesions (median 60 years, 25%/75% percentile 53.5–66 years). Figure 2 depicts representative segmentations of tumors in four patients.

The accuracy of BI-RADS MRM assessment by expert radiologist assessment using ceMRI for predicting benign and malignant lesions was 97.0% and the receiver operating
characteristic (ROC) area under the curve (AUC) 95.9%. Using the unenhanced MRM protocol, the expert radiologist achieved an accuracy of 97.0% and an ROC AUC of 95.9%.

During initial model fitting, identifiability issues in resampling techniques used for choosing the tuning parameter in Lasso regularization led us to exclude ADC\textsubscript{DWI} parameters and all ADC\textsubscript{DWIBS} parameters other than first-order parameters. In total, parameters interpreted as having a lack of relevant textural information were excluded from subsequent analysis as follows: radiomics parameters with zero variance over all observations ($n = 25$), parameters with duplicate information content ($n = 82$), parameters with less than 10 unique values after rounding to five digits ($n = 6$), and the exclusions for ADC\textsubscript{DWIBS} parameters and ADC\textsubscript{DWI} as above together with parameters with a correlation coefficient more than 0.8 with mean ADC\textsubscript{DWIBS} ($n = 287$). Diffusion data were corrupted by severe artifact in one patient, leading to exclusion of this patient from further analysis.

We fit the constrained and unconstrained L1-penalized regression models to the remaining 325 parameters. The associated tuning parameters controlling the degree of penalization (and thus the number of selected features) were estimated by 10-fold cross-validation. The resulting signature reduced the feature space to 11 radiomics parameters, each of which were included in the final model (Table 1).

As a feature importance criterion we report the feature selection frequencies over 1000 bootstrap samples.\textsuperscript{31} They provide a description of the stability of the estimated signature. A graphical depiction of the selection frequencies of selected parameters is shown in Fig. 3. The features selected into the unconstrained model included four $T_2$-weighted, two ADC, and five DWIBS\textsubscript{1500} parameters. These included the four first-order parameters, mean ADC\textsubscript{DWIBS}, minimum ADC\textsubscript{DWIBS}, and $T_2$-weighted root mean square (RMS) and mean absolute deviation. Higher-order

![FIGURE 2: Example segmentations shown on DWIBS maps (left column, top), $T_2$-weighted images (right column, top), ADC\textsubscript{DWIBS} (left column, bottom), and ADC\textsubscript{DWI} (right column, bottom) maps. Top left (A) Malignant lesion. A 67-year-old patient with a BI-RADS 5 lesion of 4 mm in the screening X-ray mammography; histopathological analysis demonstrated an invasive ductal carcinoma (IDC). Top right (B) Benign lesion. A 60-year-old patient with a BI-RADS 4 lesion on X-ray mammography, located in the left breast, size 4 mm; histopathology revealed a benign fibroadenoma. Bottom left (C) Malignant lesion. A 51-year-old patient with a BI-RADS 5 lesion in the screening X-ray mammography in the right breast; histopathological analysis revealed an invasive ductal carcinoma (IDC). Bottom right (D) Benign lesion. A 51-year-old patient with a BI-RADS 4 lesion on X-ray screening mammography located in the right breast; histopathology revealed benign fibrocystic changes.](image-url)
parameters included five co-occurrence and one run length texture parameter for the texture features calculated from DWIBS1500 and $T_2^w$, and included predominantly standard deviation parameters. Finally, one volumetric parameter, $T_2^w$ sphericity, was included. The constrained model included the same top three parameters as the unconstrained model. Five of the 11 parameters in the constrained model were different from the unconstrained model, including a replacement of the volumetric parameter sphericity with compactness and selection of some different texture parameters; however, again with a predominance of standard deviation parameters, while maintaining 6 of the 11 parameters.

For ADC, one first-order parameter, the preselected ADC mean parameter, was selected into the constrained model, while the same two first-order $T_2^w$ parameters present in the unconstrained model also entered the constrained model.

Table 2 provides the $.632 +$ bootstrap estimates of the ROC AUC, the apparent estimate of the AUC, and the leave-one-out bootstrap estimates of the AUC as well as 0.025 and 0.975 quantiles of the latter to assess uncertainty. Figure 4 provides the associated ROC curves.

The AUC of Lasso is 84.2%/85.1% for the unconstrained/constrained models, respectively. We note that the

### TABLE 1. Two Sets of 11 Radiomics Parameters Included in the Final Constrained and Unconstrained Models With Coefficients and Bootstrap Selection Frequencies (%) (Compare Fig. 2)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Coefficient</th>
<th>Selection frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2TSEn (Original): co-occ. (1) Difference Entropy SD</td>
<td>−6.964</td>
<td>0.69</td>
</tr>
<tr>
<td>ADCDWIBS (Original): First-Order Mean</td>
<td>−1039.303</td>
<td>0.6</td>
</tr>
<tr>
<td>DWIBS1n (Original): co-occ. (1) Sum Entropy SD</td>
<td>−7.87</td>
<td>0.59</td>
</tr>
<tr>
<td>ADCDWIBS (Original): First-Order Minimum</td>
<td>−136.138</td>
<td>0.56</td>
</tr>
<tr>
<td>DWIBS1n (Original): Run Length. (256) Gray Level Nonuniformity SD</td>
<td>−0.001</td>
<td>0.47</td>
</tr>
<tr>
<td>T2TSEn (Original): Volumetric Features Sphericity</td>
<td>3.446</td>
<td>0.35</td>
</tr>
<tr>
<td>DWIBS1n (Original): co-occ. (1) Sum Variance SD</td>
<td>0.009</td>
<td>0.32</td>
</tr>
<tr>
<td>T2TSEn (Original): First-Order RMS</td>
<td>−1.413</td>
<td>0.31</td>
</tr>
<tr>
<td>T2TSEn (Original): First-Order Mean absolute deviation</td>
<td>−3.704</td>
<td>0.26</td>
</tr>
<tr>
<td>DWIBS1n (Original): co-occ. (2) Inverse Difference SD</td>
<td>−6.465</td>
<td>0.2</td>
</tr>
<tr>
<td>DWIBS1n (Original): co-occ. (3) Autocorrelation Means</td>
<td>0</td>
<td>0.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Coefficient</th>
<th>Selection frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADCDWIBS (Original): First Order Mean</td>
<td>−4615.527</td>
<td>1</td>
</tr>
<tr>
<td>DWIBS1n (Original): co-occ. (1) Sum Entropy SD</td>
<td>−12.964</td>
<td>0.78</td>
</tr>
<tr>
<td>T2TSEn (Original): co-occ. (1) Difference Entropy SD</td>
<td>−3.004</td>
<td>0.47</td>
</tr>
<tr>
<td>DWIBS1n (Original): Run Length. (256) High Gray Level Run Emphasis SD</td>
<td>0.003</td>
<td>0.43</td>
</tr>
<tr>
<td>T2TSEn (Original): Volumetric Features Compactness 2</td>
<td>2.569</td>
<td>0.4</td>
</tr>
<tr>
<td>T2TSEn (Original): co-occ. (1) Cluster Shade SD</td>
<td>0</td>
<td>0.38</td>
</tr>
<tr>
<td>T2TSEn (Original): First-Order Mean absolute deviation</td>
<td>−7.653</td>
<td>0.38</td>
</tr>
<tr>
<td>DWIBS1n (Original): co-occ. (2) Difference Entropy SD</td>
<td>−0.809</td>
<td>0.38</td>
</tr>
<tr>
<td>DWIBS1n (Original): co-occ. (1) Inverse Difference SD</td>
<td>−3.769</td>
<td>0.26</td>
</tr>
<tr>
<td>DWIBS1n (Original): co-occ. (2) Inverse Difference SD</td>
<td>−2.876</td>
<td>0.2</td>
</tr>
<tr>
<td>T2TSEn (Original): First-Order RMS</td>
<td>−0.429</td>
<td>0.14</td>
</tr>
</tbody>
</table>

For definitions of radiomic parameters, see supplement of Ref. 39. co-occ. (x) = co-occurrence matrix with step width x; RMS = root mean square.
bootstrap confidence intervals of the AUCs corresponding to the radiomics signatures cover large ranges (entirely above the AUC of 50% of the null model without any covariates) due to the relatively small sample size and high correlations between features. Thus, the difference between the AUC corresponding to MR-BIRADS and the radiomics signatures is associated with a high degree of uncertainty, and hence its relevance is difficult to interpret. In Fig. 5, the clustered z-transformed parameters selected by the Lasso are shown in a color-coded heatmap (green to red color scale, with green indicating a negative and red a positive z-score). The radiomic features are clustered by similarity using the Pearson product-moment correlation coefficient; thus, the radiomic signatures in adjacent rows are similar. The first row at the bottom of Fig. 5A,B indicates histopathological status (blue: benign, red: malignant, green: undecided—according to the interquartile definition depicted in Fig. 6, see below). The third row shows classification by the mean ADC parameter, for comparison. Columns of the radiomic features are sorted accordingly, such that a clear distinction in radiomic feature expression between benign lesions predicted as benign and malignant lesions predicted as malignant is visually apparent. It can be visually assessed from the figure that radiomic signatures in incorrectly predicted lesions or in undecided cases appear less clearly distinguished. Figure 6 depicts the distribution of predicted class probabilities over all bootstrap estimates as boxplots. Each boxplot shows the interquartile range from 0.25 to 0.75. If the boxplot intersects the 0.5 line, the result is “undecided”; otherwise it is “malignant” or “benign” as shown. The color-coding indicates the abbreviated BI-RADS assessment. It can be seen that only two additional cases were missed by each model in addition to...
The one malignant case incorrectly classified as benign by the expert radiologist, while there were three classifications discordant with pathology for benign lesions for the unconstrained and one such case for the constrained model, compared to the human expert. Of the two cases incorrectly classified by the expert, the malignant lesion was equally missed by both radiomic models, while the benign lesion was found discordant for the unconstrained model but indeterminate for the constrained model. Figure 7 demonstrates cases in which mean ADC incorrectly predicted malignant lesions as benign, with equal to better performance by the radiomic models.

**DISCUSSION**

The discrimination ability between malignant and benign lesions by the radiomic models was better than the monoparameter mean ADC. Including the mean ADC parameter as a mandatory component into the full parameter set led to an additional increase in performance, highlighting the importance of incorporating available prior knowledge into model building. Importantly, mean ADC was selected second most often into the unconstrained model, indicating that unconstrained feature selection was able to attribute a high importance to this parameter. The first-order features selected into the unconstrained model evaluate the intraleisonal ADC value and heterogeneity in the $T_{2w}$ regions, accounting for known MR characteristics important for deciding if a lesion is benign or malignant. The selection of predominantly higher-order texture standard deviation parameters (co-occurrence and run length parameters) for DWIBS1500 and $T_{2w}$ indicates that texture heterogeneity in different spatial directions was descriptive of lesion classification. The inclusion of volumetric parameters (sphericity and compactness) into the models is in accordance with the clinical observation that malignant lesions tend to assume more spherical and thus compact shapes than benign lesions. The constrained model included the same top three parameters and in total 6 of the 11 parameters of the unconstrained model, with the remaining changes occurring predominantly within related standard deviation texture features. We interpret this finding as indicative of stability in the model building process under the perturbation of introducing constraints. Preselection of mean ADC into the constrained model prevented selection of other first-order ADC parameters, confirming its importance.

None of the quantitative models reached the almost perfect performance of the highly experienced expert.
radiologist, who performed almost equally well using the unenhanced MRM protocol and standard contrast-enhanced MRM. We note that the expert performance shown here is likely above average of all screening certified radiologists and that approximating or even matching this performance would be a great success for radiomics, taking into account that such a radiomics system would provide training-independent decision support. The excellent performance of a highly trained screening radiologist is the result of long and intensive training (on a much larger number of cases than the number used to train the radiomics classifiers here) and the decision whether a lesion is graded as benign or malignant can usually not be summarized quantitatively in terms of how much information was gained from each imaging sequence. On the other hand, simple imaging metrics such as signal intensity and lesion size or visually

FIGURE 5: Heatmaps depicting z-score transformed selected radiomics features (rows) for all 49 patients (columns). Heatmaps are given (A) for the unconstrained model and (B) for the constrained model with mandatory inclusion of mean ADC. The five rows of the bottom annotation are: first row: Histologically proven malignant (red) and benign (blue) breast lesions; second row: Predictions based on bootstrap quartiles as given in Fig. 6 (“malignant” when 75% of the predicted probabilities of malignancy over 1000 bootstrap samples were above the threshold 0.5, “benign” when 75% of the predicted probabilities of malignancy over the bootstrap samples were below 0.5, and “undecided” when the 25% and 75% percentile over the bootstrap samples includes 0.5); third row: Predictions by the monoparameter ADC, color scale analogous to second row. It can be seen that ADC overcalls more benign cases as malignant compared to Lasso (left). Compare Fig. 7 for cases in which ADC incorrectly predicted a malignant lesion to be benign; fourth and fifth rows: for comparison, BI-RADS (third row) and abbreviated BI-RADS (fourth row) scoring of lesions is shown according to the legend. For definitions of radiomic parameters, see Data Supplement S1. BI-RADS = Breast Imaging Reporting and Data System; T2TSE = T2-weighted turbo-spin echo; DWIBS = diffusion-weighted imaging with background suppression.
assessed degree of lesion conspicuity (eg, based on spiculation) do not describe or exploit all of the information that is captured in the images. Radiomics has recently been developed to assess a large number of imaging features to characterize the tumor phenotype to predict clinical outcomes with increased accuracy. We point out that the analysis performed here provides a context in which model performance can be assessed under current conditions; however, this does not exclude the expert’s performance as a limit that can be achieved with extension of the technique. The .632 + bootstrap analysis is conservative to avoid frank over-optimistic assessment of performance and provides a critical estimate that the examined models appear in the described order under the examined data. Taking into account the distribution of the 2.5% and 97.5% quantiles of ROC AUC, the model fit still appears rather unstable in a limited dataset of this size; however, we hypothesize that model fitting in larger datasets will lead to stabilization and allow approximation of a higher performance level. It is important that the results presented here are compared to equally conservative estimates of model performance and to an equally demanding gold standard used in other radiomics studies. Specifically, we point out that the apparent accuracy estimates would have been up to 99.4% for our model. Apparent estimates are derived from estimating the performance of the model on the training set. Clearly, such a validation strategy would lead to excessive overestimation of the classification performance and cannot be recommended. We note that clear incorrect decisions were made by the unconstrained/constrained models in only three/one benign and three/three malignant lesions compared to one in each category for the expert radiologist. With this performance an added value of the current models as a clinical decision tool is suggested in certain situations. The “undecided” cases show that most misclassifications might have occurred when Lasso predictions were highly variable. This additionally supports the prospect that a more stable radiomic signature based on a larger sample size is anticipated to provide a better prediction performance with reduction of indeterminate assessments. To our knowledge, this is the first study

![Boxplots depicting predicted probabilities of malignancy in 1000 bootstrap samples for all lesions in the sample (A) for the unconstrained model and (B) for the constrained model with mandatory inclusion of mean ADC. Histologically benign lesions are shown to the left of the vertical line, histologically malignant lesions to the right. The ends of the vertical boxplots indicate the first and third quartiles, their length the interquartile range, and the horizontal black line the median of classification probability as malignant of the respective lesions by the radiomics classifiers. Boxplots that intersect the horizontal line were considered undecided, those lying entirely above (below) the horizontal line as a prediction of malignant (benign). Colors indicate abbreviated BI-RADS scores and provide a visual comparison between human expert and radiomics classifier. The horizontal line corresponds to a predicted probability by the radiomic model for the presence of malignancy of 0.5.](image-url)
examining the potential of radiomics in unenhanced MRM with DWIBS sequences in the highly relevant field of breast cancer screening programs.

Bogner et al reported an accuracy in the distinction of benign and malignant lesions by ADC measurement of 95% in 51 patients with 112 ROIs in benign, cystic, normal, and malignant tissue and determined that a combined b-value protocol of 50 and 850 s/mm² had the best imaging quality at 3.0T. We found a lower performance of ADC in our study, which may be attributed to the preselected cohort of patients having mammographically suspicious lesions. We included patients with BI-RADS 4 and 5

FIGURE 7: Overview of malignant lesions misclassified by ADC. ROIs (red) shown overlaid on DWIBS maps (left column, top), T₂-weighted images (right column, top), ADCDWI (left column, bottom), and ADCDWIBS (right column, bottom) maps. Top left (A) Classified by mean ADC as indeterminate, by both constrained and unconstrained Lasso models as malignant. A 65-year-old patient with BI-RADS 5 lesion in the screening X-ray mammography revealing an adenosquamous invasive ductal carcinoma (IDC) in histopathology. Top right (B) Classified as benign by mean ADC, as indeterminate by unconstrained and as benign by constrained Lasso model. A 53-year-old patient with a BI-RADS 4 lesion on X-ray mammography, revealing an invasive ductal carcinoma (IDC) in histopathology. Bottom left (C) Classified as benign by mean ADC, as malignant by unconstrained, and as indeterminate by constrained Lasso models. A 57-year-old patient with a BI-RADS 4b lesion on screening X-ray mammography revealing an invasive ductal carcinoma (IDC) with mainly intraductal components in histopathology. Bottom right (D) Classified as indeterminate by all models (mean ADC, constrained and unconstrained Lasso models). A 68-year-old patient with a BI-RADS 4b lesion on screening X-ray mammography revealing with an invasive ductal carcinoma (IDC) in histopathology misclassified using ADC.
lesions on mammography, while the mentioned study included BI-RADS 0, 3–5 lesions (with 22 BI-RADS 0 and 4 BI-RADS 3 lesions, and only 8 BI-RADS 4 and 17 BI-RADS 5 lesions). The b-value scheme used here is the same that was available to the expert radiologist, who achieved excellent performance with it. For comparability, we did not change the b-value settings for the radiomics analysis presented here.

Breast cancer screening programs have repeatedly been challenged due to the relatively high rate of false-positive findings causing invasive procedures in healthy participants.33,34 Methods to overcome this problem have been investigated and DWI has recently been identified as a promising sequence to stratify suspicious breast lesions with unenhanced MRI11,12 with the opportunity to decrease costs at an acquisition time of less than 7 minutes.31 Although there have been reports that when used concomitantly with DCE-MRI, DWI may not offer an added diagnostic advantage over DCE-MRI,35 in view of the recent increased awareness of deposition of gadolinium in the brain36,37 the development of contrast-free examinations even of comparable accuracy appears to be highly attractive. The combination of breast radiomics with unenhanced breast MRI protocols is attractive, as it may develop into a tool to decrease user-dependence of interpretation. Currently, only well-trained and certified experts are working in breast cancer screening, since it is known that diagnostic accuracy highly depends on the experience of the reader.

Limitations of our study include the small cohort size which precluded us from splitting the cohort into a training and validation set. A single such split would have lowered the size of both sets too much to allow for a useful analysis. We applied .632 + bootstrapping to evaluate the performance of many such splits and report the average performance. The prospective use of the classifier in an extended cohort will be subject to further study. There is a certain selection bias towards higher-risk disease in the lesions analyzed here compared to all questionable lesions that are detected on MRM, as all examined lesions had a suspicious mammography correlate. Manual lesion segmentation is operator-dependent, and automated and semiautomated approaches have been proposed, eg, the method of image manifold revealing the segmentation of DCE-MRI data on MRM examinations.38 However, given the small study cohort careful manual lesion segmentation was feasible and matched the lesions identified by expert readers. For larger cohorts the time demand for manual segmentation may be prohibitive, and automated segmentation algorithms are required to be developed for success in analyzing very large datasets. Finally, breast lesions are often of small volume and we did not apply a size threshold. It is possible that the usefulness of radiomics may be limited in very small lesions in which only few voxel data are available for textural analysis. While this may have reduced the performance of our models compared to radiomics analyses focusing only on larger lesions, we consider it important to develop analysis techniques that can extract useful information from radiomics data in all clinically applicable settings and can handle any data that are accessible to human readers to allow a direct and unbiased comparison of the ability of radiomics to extract diagnostically important information from image data.

In conclusion, in this pilot study we identified two radiomics classifiers from a contrast-free abbreviated MRM protocol that performed well in the differentiation of malignant from benign lesions and achieved higher performance than the mean ADC parameter alone. Mandatory inclusion of mean ADC into the model led to an additional performance increase. Classification was lower than the almost perfect performance of a highly experienced breast radiologist using either standard ceMRM and unenhanced DWIBS-based MRM. The potential of radiomics in predictive models of breast cancer imaging to provide a training-independent diagnostic decision tool in the future is indicated. A performance reaching the human expert would be highly desirable and is considered possible from our results when the concept is extended in larger cohorts with further development and validation of the technique. With additional validation, texture analysis may become a useful non-invasive addition to clinical assessment in the future without additional harm and cost.

ACKNOWLEDGMENT
Contract grant sponsor: Dietmar Hopp Foundation
P.K. is a fellow of the Medical Faculty Heidelberg Postdoc-Program.

AUTHOR CONTRIBUTIONS
S.B., D.B., and H.P.S. conceived and designed the study; S.B., D.P., and F.S. performed image postprocessing; D.B., D.T., P.K., and M.W. performed statistical analysis; K.M.H., N.G. and M.G. developed customized imaging and radiomics software; F.B.L. supervised medical physics optimization of the diffusion protocol; S.B., W.L., and H.D. were involved in patient acquisition and clinical follow-up; W.L. and H.D. performed and supervised clinical image interpretation; D.B., S.B., P.K., and D.P. wrote large parts of the article; K.M.H., M.G., M.W., D.T., F.S., W.L., H.D., N.G., F.B.L. and H.P.S. contributed to the article.

CONFLICT OF INTEREST
There are no competing financial interests in relation to the work described.
REFERENCES


学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，
提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。
图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：

图书馆首页 文献云下载 图书馆入口 外文数据库大全 疑难文献辅助工具