Eur Radiol (2016) 26:2529–2537
DOI 10.1007/s00330-015-4075-7

BREAST

A simple scoring system for breast MRI interpretation: does it compensate for reader experience?

Maria Adele Marino[1,2] · Paola Clauser[1,3] · Ramona Woitek[1] · Georg J. Wengert[1] · Panagiotis Kapetas[1] · Maria Bernathova[1] · Katja Pinker-Domenig[1] · Thomas H. Helbich[1] · Klaus Preidler[4] · Pascal A. T. Baltzer[1]

Received: 16 July 2015 / Revised: 12 October 2015 / Accepted: 16 October 2015 / Published online: 29 October 2015

Abstract
Purpose To investigate the impact of a scoring system (*Tree*) on inter-reader agreement and diagnostic performance in breast MRI reading.
Materials and methods This IRB-approved, single-centre study included 100 patients with 121 consecutive histopathologically verified lesions (52 malignant, 68 benign). Four breast radiologists with different levels of MRI experience and blinded to histopathology retrospectively evaluated all examinations. Readers independently applied two methods to classify breast lesions: BI-RADS and *Tree*. BI-RADS provides a reporting lexicon that is empirically translated into likelihoods of malignancy; *Tree* is a scoring system that results in a diagnostic category. Readings were compared by ROC analysis and kappa statistics.
Results Inter-reader agreement was substantial to almost perfect (kappa: 0.643–0.896) for *Tree* and moderate (kappa: 0.455–0.657) for BI-RADS. Diagnostic performance using *Tree* (AUC: 0.889–0.943) was similar to BI-RADS (AUC: 0.872–0.953). Less experienced radiologists achieved AUC: improvements up to 4.7 % using *Tree* (*P*-values: 0.042–0.698); an expert's performance did not change (*P*=0.526). The least experienced reader improved in specificity using *Tree* (16 %, *P*=0.001). No further sensitivity and specificity differences were found (*P*>0.1).
Conclusion The *Tree* scoring system improves inter-reader agreement and achieves a diagnostic performance similar to that of BI-RADS. Less experienced radiologists, in particular, benefit from *Tree*.
Key Points
- *The Tree scoring system shows high diagnostic accuracy in mass and non-mass lesions.*
- *The Tree scoring system reduces inter-reader variability related to reader experience.*
- *The Tree scoring system improves diagnostic accuracy in non-expert readers.*

Keywords Breast cancer · MRI · Scoring system · Reader experience · Sensitivity and specificity

✉ Pascal A. T. Baltzer
pascal.baltzer@meduniwien.ac.at

[1] Department of Biomedical Imaging and Image-guided Therapy, Division of Molecular and Gender Imaging, Medical University of Vienna, Vienna General Hospital, Floor 7F Waehringer Guertel 18-20, 1090 Vienna, Austria

[2] Department of Biomedical Sciences and Morphologic and Functional Imaging, Policlinico Universitario G. Martino, University of Messina, Messina, Italy

[3] Department of Medical and Biological Sciences, Institute of Diagnostic Radiology, Azienda Ospedaliero-Universitaria, "S. Maria della Misericordia", P.le Santa Maria della Misericordia, University of Udine, Udine, Italy

[4] Diagnosezentrum Meidling, Vienna, Austria

Introduction

Magnetic resonance imaging of the breast has been established worldwide as a highly accurate imaging modality for the detection of breast cancer, with an excellent sensitivity up to 100 % [1–3]. However, breast MRI can be challenging: many different criteria can be used for image interpretation, and technical recommendations encompass a broad variety of examination and interpretation quality. Several recommendations have been published that were designed to improve the standardization of breast MRI acquisition and reporting [4–8]. The most widely accepted standard is the American College of Radiology (ACR) Breast Imaging Reporting and Data System

(BI-RADS) lexicon [7]. It contains a structured common language for interpretation and reporting of mammography (MG), ultrasound (US), and MRI. Without a doubt, the BI-RADS lexicon facilitates communication among physicians through the use of a standardized terminology. The MRI BI-RADS lexicon features cover lesion morphology, such as margins, internal enhancement pattern, and functional contrast enhancement kinetics. However, the BI-RADS lexicon does not provide defined rules by which to convert specific imaging features into a diagnostic category [7]. Moreover, the use of multiple diagnostic criteria is associated with the risk of information redundancy [9]. As a consequence, inter-reader agreement of BI-RADS is generally moderate while diagnostic accuracy is highly variable [10–13].

A scoring system is defined as a clinical decision rule that leads to a prognostic estimation or a diagnostic category by incorporating several criteria [14]. Scoring systems have been investigated in order to assist radiologists in characterising MRI findings and improving their specificity for the prediction of breast lesion malignancy [8, 15–21]. Baltzer et al. proposed a classification *Tree* flowchart as a structured and intuitive algorithm for the differentiation of malignant and benign lesions [15]. In that algorithm, five diagnostic criteria independently contribute to lesion diagnosis and each specific combination of criteria provides a likelihood of malignancy. Such a scoring system would be expected to improve inter-reader agreement and may reduce experience-related variability. According to the Centre of Evidence-Based Medicine reasoning, the *Tree* scoring system of Baltzer et al. was exploratory. In order to achieve a high level of diagnostic accuracy, a validating study in one clinical centre is required [14].

Our aim was to investigate the diagnostic performance and inter-reader agreement of the *Tree* scoring system and compare these parameters with standard MRI BI-RADS lexicon reading.

Materials and methods

Study cohort

Our institutional review board (IRB) approved this retrospective diagnostic single-centre study and waived the necessity for informed consent. Four hundred and fifty-nine consecutive patients undergoing MRI examinations from April 4th, 2013, to September 25th, 2014, were eligible. MRI, in accordance with international recommendations, was performed to evaluate the following conditions: *(a)* unclear findings (conventional imaging BI-RADS 0 and 3); *(b)* suspicious lesions or lesions highly suggestive of malignancy (conventional imaging BI-RADS 4 and 5); and *(c)* preoperative staging of biopsy–proven breast cancer (BI-RADS 6). No high-risk screening patients were included in this study. Further inclusion criterion was a reference standard by means of histopathological analysis, either by image-guided biopsy (US-guided core biopsy or vacuum-assisted biopsy under MG/MRI guidance) or open surgery according to international guidelines for quality assurance in breast cancer screening and diagnosis [22]. Board-certified breast pathologists performed the work-up of breast tissue specimens. Patient selection details are given in Fig. 1.

Magnetic resonance imaging of the breast

MRI examinations were performed on a 1.5 T (Siemens Espree; Erlangen, Germany) with the patient in the prone position, using a dedicated bilateral four-channel breast coil. The examination protocol included: an axial T2-weighted turbo spin echo (TSE) without fat saturation (TR/TE: 5160/197 ms; spatial resolution 0.7×0.7 × 3 mm, acquisition time 3:07 min.); an axial turbo inversion recovery magnitude sequence (TIRM) (TR/TE 10460/64; spatial resolution 1.1×1.1×4 mm, acquisition time 2:59 min.); an axial diffusion-weighted echo-planar imaging (DW-EPI) sequence (TR/TE 6300/117 ms, spatial resolution 1.6×1.6×3 mm, b values of 0 and 1000 s/mm^2, fat saturation by water excitation, 12 averages, acquisition time 2:50 min.). During the time course of this study, two different dynamic contrast-enhanced protocols were used because the protocol was modernized. Contrast agent dosage and injection (20 mL Gadoteridol [ProHance, BRACCO, Italy] 2.5 mL/s, 20 mL saline solution) were kept constant. Protocol 1 implemented an axial T1-weighted 2D-FLASH sequence (TR/TE: 129/4.74 ms; spatial resolution 0.9×0.9×3 mm, acquisition time per measurement 1:09 min., one measurement before and four after contrast medium injection) and the modernized protocol 2 used an axial T1-weighted 3D-FLASH sequence with radial k-space sampling and Dixon fat saturation (TR/TE$_1$/TE$_2$: 12.4/4.77/9.54; spatial resolution 0.7×0.7×2 mm, acquisition time per measurement 1:35 min., one measurement before and three after contrast medium injection). ADC maps and scaled subtractions were calculated automatically, using the vendor-supplied scanner software. All image data sets were stored in our picture archiving and communication system (PACS, IMPAX EE, AGFA, Bonn, Germany).

Data analysis

All examinations were independently analysed by four breast imaging radiologists, blinded to the initial radiological BI-RADS (both at conventional and MR imaging) category assignment and the final histopathological diagnosis. The readers had different levels of experience in breast MRI and they were classified according to the number of cases read prior to this study with histological verification: less experienced (R1, 200

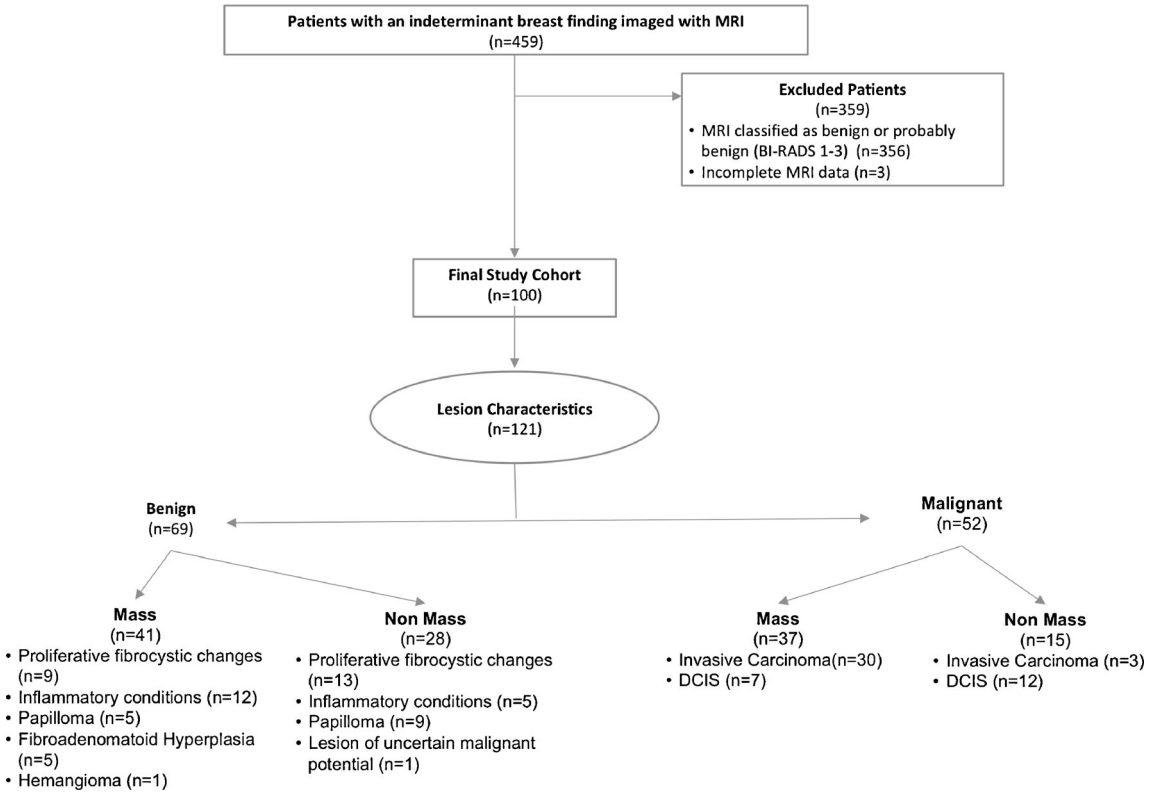

Fig. 1 Patient selection flow-chart and final lesion diagnoses stratified by presentation as mass or non-mass

cases); intermediately experienced (R2, R3, 600 cases each); and highly experienced (R4, >5000 cases).

They were asked to assign a BI-RADS rating to each lesion according to the suspicion of malignancy (BI-RADS 2 to BI-RADS 5), in congruence with the fifth edition of the MRI BI-RADS lexicon [7]. The results were stored in a spreadsheet. Subsequently, the radiologists followed a previously published scoring system (*Tree*) based on five diagnostic criteria and containing 11 assignment categories that corresponded to an increasing probability of malignancy (1 = lowest, cancer very unlikely, to 11 = highest, cancer highly probable; Fig. 2, [15]). A detailed explanation of the diagnostic criteria of the *Tree* including schematic drawings and example cases is provided as Supplemental Material 1. A diagnostic category was chosen by following the *Tree* criteria and noted in a spreadsheet. Before the beginning of image analysis, a training session was held, in which readers were shown 10 example cases that were not part of the study cohort, to demonstrate the application of the *Tree* scoring system. Results for mass and non-mass lesions are given in Figs. 3 and 4.

Statistical analysis

Statistical analyses were performed using SPSS 22.0 (SPSS, IBM, USA) and MedCalc 15 (MedCalc software bvba, Ostend, Belgium). All calculations were primarily performed on a per-lesion basis. Final diagnosis, as the reference standard for each lesion, was obtained from histopathology reports prospectively stored in our institutional database and prospectively checked for congruence during our weekly interdisciplinary meetings. Inter-reader agreement in the assigned MRI BI-RADS and *Tree* categories was assessed using kappa statistics. A receiver operating characteristic (ROC) analysis was performed to determine overall diagnostic performance, measured by the area under the ROC curve. Further, sensitivity and specificity and likelihood ratios were calculated at cut-off values of BI-RADS >3 and *Tree* >4 and >7. Statistical significance was assumed at *P*-values ≤ 0.05.

Results

Lesion characteristics

Our final study cohort consisted of 100 subjects (mean age 53±14 years [standard deviation]; age range 25–88 years). Here, a total of 121 lesions were histologically verified, 52 (43 %) of them malignant and 69 (57 %) benign. Thirty-seven malignant lesions presented as

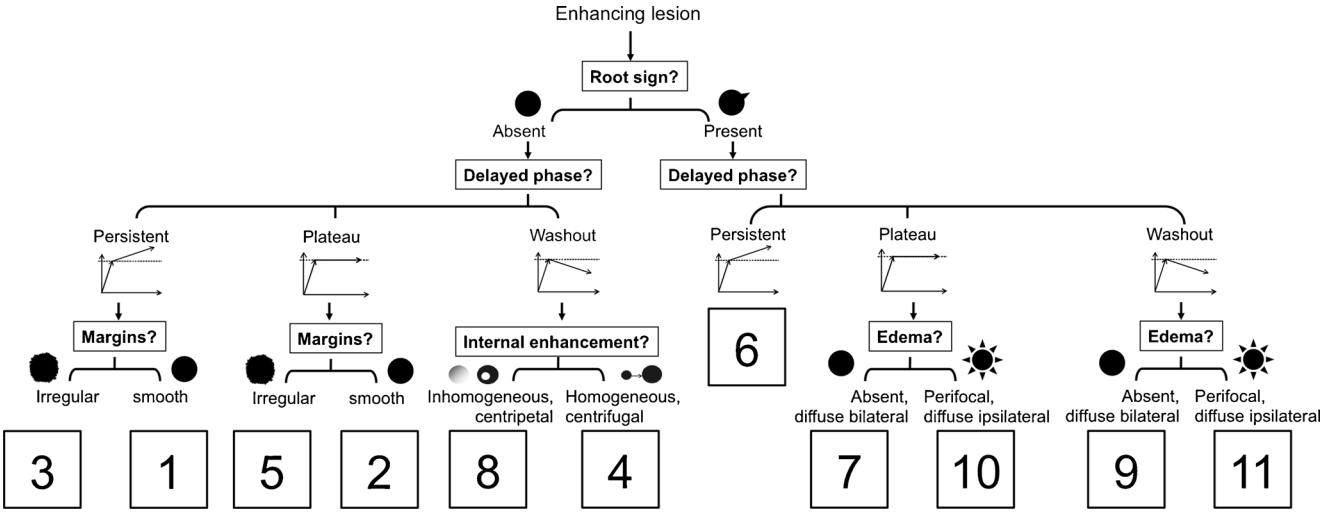

Fig. 2 *Tree* scoring system flow-chart following the initial description by Baltzer et al. 2014 [15]. Terminal nodes are hierarchically ordered and denote an increasing probability of malignancy (1 = lowest, 11 = highest)

masses (71 %; mean size 2±1.1 cm; range 0.7–4.9 cm), 15 presented as non-mass lesions (29 %; mean size 4.16± 3.3 cm; range 1–13 cm). Forty-one (60 %) benign lesions were masses (mean size 1.39±0.9 cm; range 0.5–5 cm) and 28 (40 %) presented as non-mass lesions (mean size 2 ±2.5 cm; range 0.6–10 cm).

Details on lesion diagnoses stratified by presentation as mass or non-mass are given in Fig. 1.

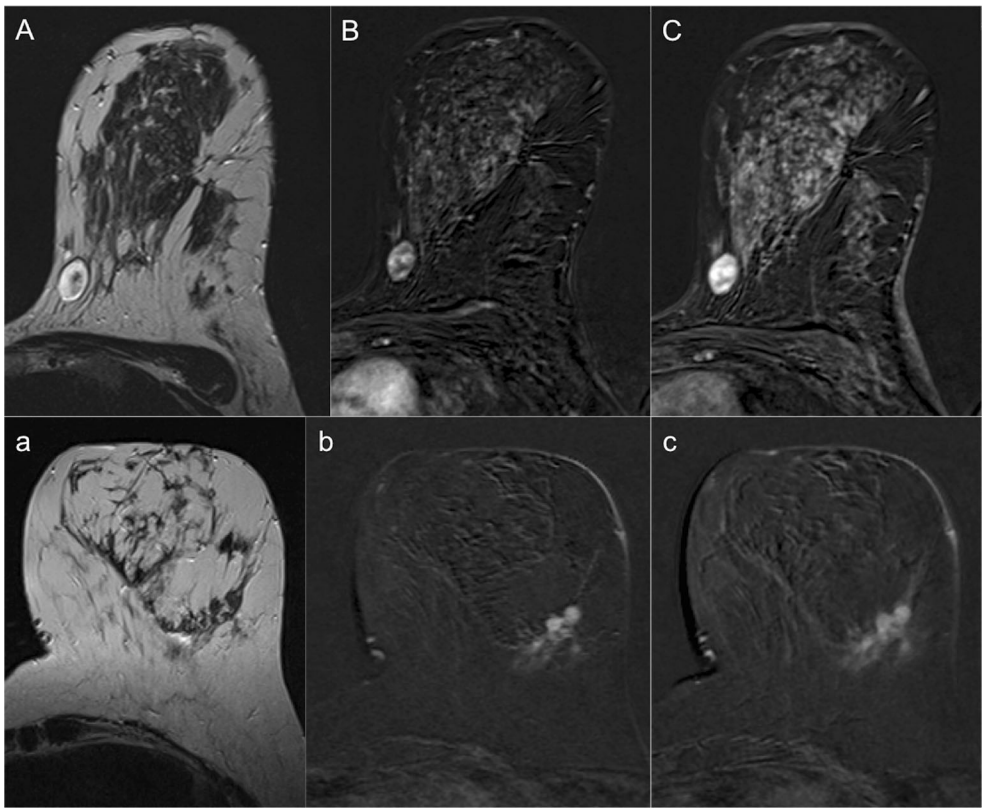

Fig. 3 Mass lesion examples. Upper row (*capital letters*) shows a fibroadenoma, presenting as a mass lesion with circumscribed margins, heterogeneous internal enhancement, and a persistent signal enhancement time curve. Based on the *Tree* system (Fig. 2), the absence of the root sign resulted in a node 1 (benign finding most likely) rating. The lesion was classified as BI-RADS 3 (T2w *A*, early *B*, and late *C* subtractions). The lower row shows an invasive carcinoma presenting as an irregularly shaped mass lesion. Margins are not circumscribed, with some small spiculations. The curve type is washout and a perifocal oedema is present as high SI on T2w. Based on the *Tree* system, the root sign is present, the curve type is washout, and perifocal oedema is present, which resulted in a node 11 (malignant finding most likely) rating. T2w (a), early (b), and late (c) subtractions

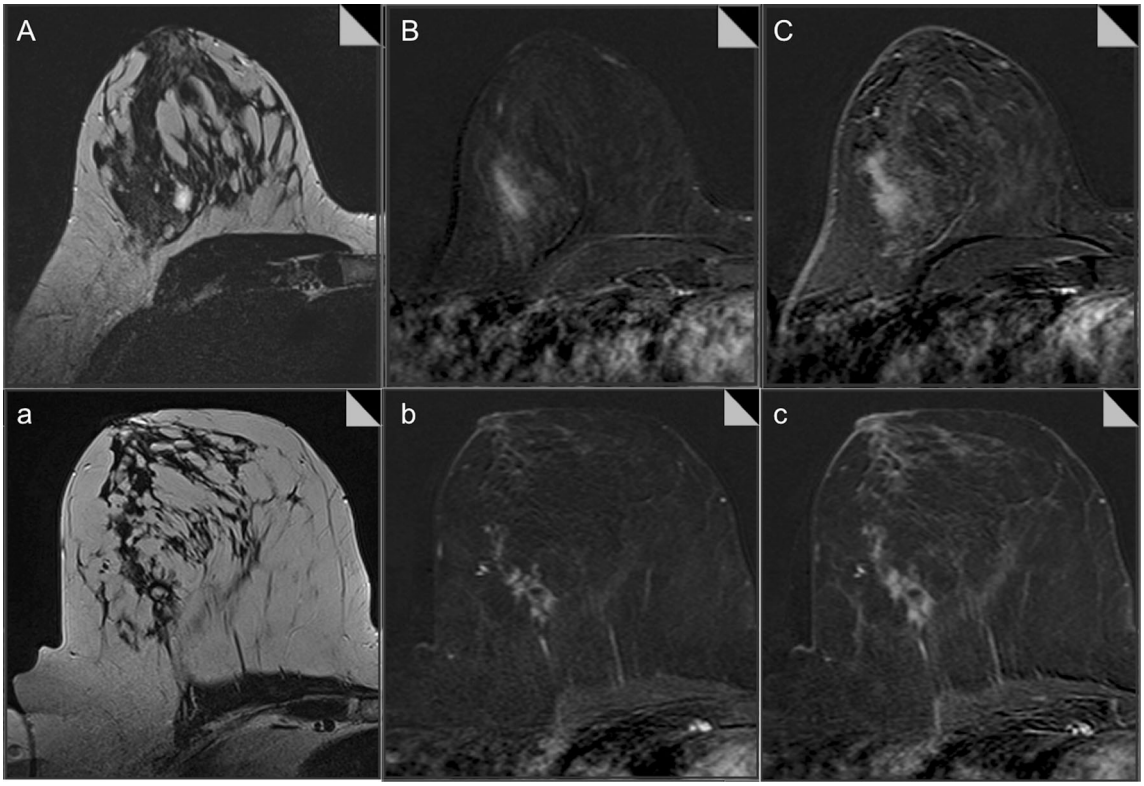

Fig. 4 Non-mass lesion examples. Upper row (*capital letters*) shows a focal inflammation, presenting as a regional non-mass lesion with homogeneous enhancement and a persistent signal enhancement time curve. Based on the *Tree* system (Fig. 2), the absence of a root sign, the persistent time curve type, and the non-circumscribed margins resulted in a node 3 (benign finding likely) rating. The lesion was classified as BI-RADS 3 (T2w *A*, early *B*, and late *C* subtractions). The lower row shows an invasive lobular carcinoma presenting as a clumped segmental non-mass lesion. Some readers reported persistent signal enhancement and some readers reported plateau enhancement time curves. Based on the *Tree* system, the root sign is present, the curve type is either persistent or plateau, and oedema is absent, resulting in a node 6 or 7 rating (malignant finding suspected, refer to biopsy). T2w (a), early (b), and late (c) subtractions

Inter-reader agreement

The kappa agreement among the four readers for the characterization of breast lesions as cancers was substantial to almost perfect (k=0.643–0.896) for *Tree*, while it was only moderate for BI-RADS (k=0.455–0.657, Table 1).

Area under the ROC curve

The area under the ROC curve (AUC) for lesion diagnosis ranged between 0.889 and 0.943 for *Tree* 0.872 and 0.953 for BI-RADS (Fig. 5). The AUC of *Tree* reading was higher than BI-RADS in the less-experienced and intermediately-

Table 1 Kappa values and AUC difference significance level for BI-RADS and *Tree* in all four readers

	BI-RADS R1	BI-RADS R2	BI-RADS R3	BI-RADS R4	*Tree* R1	*Tree* R2	*Tree* R3	*Tree* R4
BI-RADS R1	-	0.455^{+}	0.527^{+}	0.535^{+}	0.497^{+}	0.542^{+}	0.527^{+}	0.513^{+}
BI-RADS R2	0.686	-	0.482^{+}	0.486^{+}	0.564^{+}	0.636^{++}	0.490^{+}	0.604^{++}
BI-RADS R3	0.825	0.849	-	0.657^{++}	0.529^{+}	0.567^{+}	0.666^{++}	0.639^{++}
BI-RADS R4	0.0004*	0.012*	0.0024*	-	0.568^{+}	0.570^{+}	0.635^{++}	0.607^{++}
Tree R1	0.153	0.263	0.261	0.047*	-	0.860^{+++}	0.643^{++}	0.896^{+++}
Tree R2	0.046*	0.042*	0.081	0.334	0.388	-	0.647^{++}	0.896^{+++}
Tree R3	0.631	0.902	0.698	0.035*	0.431	0.281	-	0.720^{++}
Tree R4	0.009*	0.165	0.022*	0.526	0.026*	0.489	0.086	-

UPPER RIGHT TRIANGLE: kappa agreement (kappa interpretation: 0.41–0.60=moderate $^{+}$, 0.61–0.80=substantial $^{++}$, 0.81–1=almost perfect agreement $^{+++}$), LOWER LEFT TRIANGLE: *P*-values for AUC differences, significant *P*-values indicated by * (see Table 1 for AUC values)

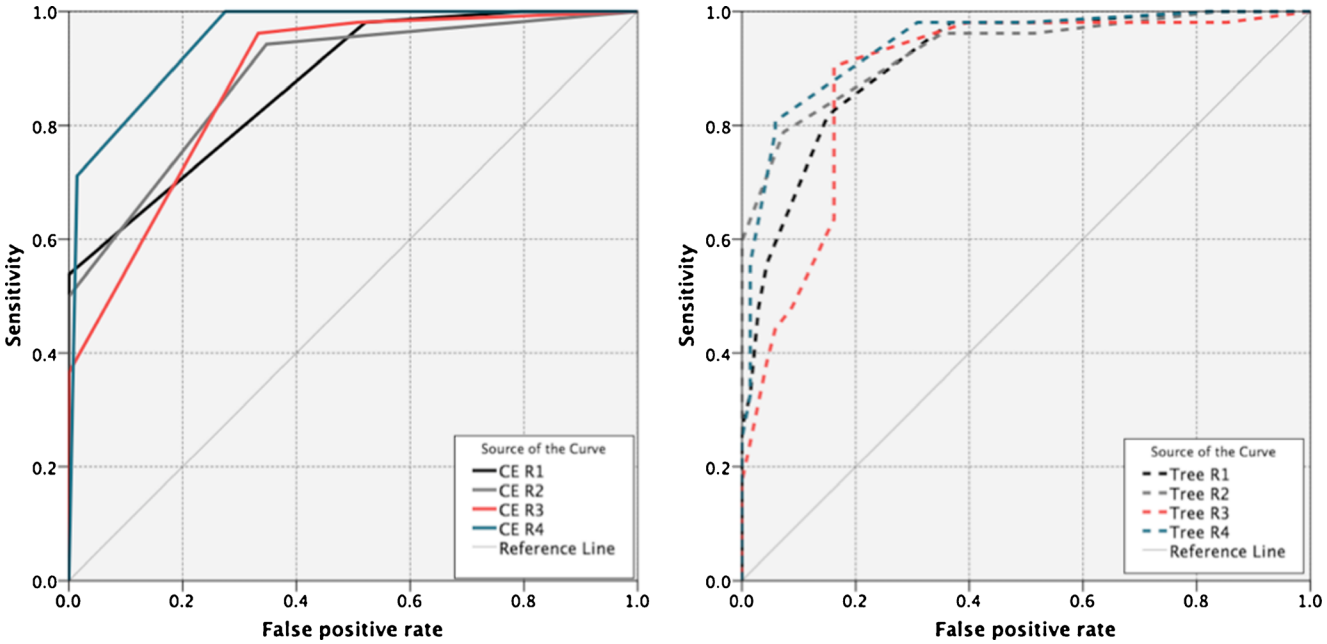

Fig. 5 Receiver operating characteristics (ROC) curves of BI-RADS (left) and *Tree* (right) ratings. Note that a high sensitivity level, corresponding to a rule-out criterion, was achieved for all observers using both techniques, but with a more uniform specificity for Tree (for details see Tables 1 and 3)

experienced readers; however, this was only significant for R2 (Tables 1 and 2).

No significant differences were found between the AUC of intermediately experienced readers (R2, R3) and the expert reader (R2/R4 P=0.4; R3/R4 P=0.08) in *Tree* reading, but there was a significant difference between R1 and R4 (P=0.02). In BI-RADS reading, the expert reader (R4) performed significantly better than all less-experienced readers (R1/R4 P=0.004; R2/R3 P=0.01 R3/R4 P=0.002).

Sensitivity and specificity

Detailed reading results using BI-RADS and *Tree*, with the corresponding sensitivity and specificity values and 95 % confidence intervals for each reader stratified by presentation as mass or non-mass are displayed in Table 3.

Tree achieved sensitivity equal to that of BI-RADS readings (differences in sensitivity each P>0.1), ranging between 96.2 and 98.1 % (Table 3). Specificity was significantly improved by *Tree* reading in the inexperienced reader (R1, specificity difference: 16 %, 95 %CI 6.9–16 %, P=0.001). False-positive cases were reduced by 30.6 % (36 FP with BI-RADS vs. 25 FP with *Tree*). Specificity did not change in the readers with intermediate and high experience (P>0.1). Similar results were observed in mass and non-mass lesions. R1 had a higher specificity in mass lesions using *Tree* compared to BI-RADS [difference 17.1 % (95 %CI 3.1–17.1 %), P=0.0156]. No further differences were observed between *Tree* and BI-RADS in mass and non-mass lesions (P>0.1, respectively). The sensitivity and specificity of *Tree* did not differ between masses and non-masses (P>0.1, respectively).

Table 2 Areas under the ROC curves (AUC) for BI-RADS and *Tree* reading for all four readers with their corresponding standard errors and 95 % confidence intervals

Test Result Variable(s)	Area	95 % Confidence Interval	
		Lower Bound	Upper Bound
BI-RADS R1	0.872	0.811	0.933
BI-RADS R2	0.884	0.824	0.944
BI-RADS R3	0.878	0.819	0.937
BI-RADS R4	0.953	0.920	0.986
Tree R1	0.914	0.865	0.963
Tree R2	0.931	0.885	0.977
Tree R3	0.889	0.828	0.950
Tree R4	0.943	0.902	0.983

Discussion

Our study investigated the inter-reader agreement and diagnostic performance of *Tree*, a scoring system for breast MRI. *Tree* leads the radiologist step-by-step toward the final diagnosis of a breast lesion detected on MRI, using an intuitive flow-chart that is easy to follow [15]. We found that *Tree* is a highly accurate scoring system, improving inter-reader agreement and achieving a high diagnostic accuracy. This was especially evident for the least experienced radiologist. These results have important clinical implications: in addition to BI-

Table 3 Diagnostic parameters of BI-RADS reading results and *Tree* reading results for all four readers stratified by lesion presentation as mass or non-mass lesions

Criterion	Sensitivity (TP/TP+FN)	95 % CI	Specificity (TN/TN+FP)	95 % CI	+LR	-LR
All lesions						
BI-RADS R1 >3	98.1 (51/52)	89.7 - 100	47.8 (33/69)	35.6 - 60.2	1.88	0.040
BI-RADS R2 >3	94.2 (49/52)	84.1 - 98.8	65.2 (45/69)	52.8 - 76.3	2.71	0.088
BI-RADS R3 >3	96.2 (50/52)	86.8 - 99.5	66.7 (46/69)	54.3 - 77.6	2.88	0.058
BI-RADS R4 >3	100 (52/52)	93.2 - 100	72.5 (50/69)	60.4 - 82.5	3.63	<0.01
Tree R1 >4	98.1 (51/52)	89.7 - 100	63.8 (44/69)	51.3 - 75.0	2.71	0.030
Tree R2 >4	96.2 (50/52)	86.8 - 99.5	65.2 (45/69)	52.8 - 76.3	2.76	0.059
Tree R3 >4	98.1 (51/52)	89.7 - 100	60.9 (42/69)	48.4 - 72.4	2.51	0.032
Tree R4 >4	98.1 (51/52)	89.7 - 100	69.6 (48/69)	57.3 - 80.1	3.22	0.028
Mass lesions						
BI-RADS R1 >3	100 (37/37)	90.5 - 100	47.8 (20/41)	32.9 - 64.9	1.95	<0.01
BI-RADS R2 >3	97.3 (36/37)	85.8 - 99.9	65.9 (27/41)	49.4 - 79.9	2.85	0.041
BI-RADS R3 >3	97.3 (36/37)	85.8 - 99.9	65.9 (27/41)	49.4 - 79.9	2.85	0.041
BI-RADS R4 >3	100 (37/37)	90.5 - 100	78.1 (32/41)	62.4 - 89.4	4.56	<0.01
Tree R1 >4	97.3 (36/37)	85.8 - 99.9	65.9 (27/41)	49.4 - 79.9	2.85	0.041
Tree R2 >4	94.6 (35/37)	81.8 - 99.3	70.7 (29/41)	54.5 - 83.9	3.23	0.076
Tree R3 >4	97.3 (36/37)	85.8 - 99.9	61.0 (25/41)	44.5 - 75.8	2.49	0.044
Tree R4 >4	97.3 (36/37)	85.8 - 99.9	70.7 (29/41)	54.5 - 83.9	3.32	0.038
Non-mass lesions						
BI-RADS R1 >3	93.3 (14/15)	68.1 - 99.8	46.4 (13/28)	27.5 – 66.1	1.74	0.14
BI-RADS R2 >3	86.7 (13/15)	59.5 - 98.3	64.3 (18/28)	59.5 - 98.3	2.43	0.21
BI-RADS R3 >3	93.3 (14/15)	68.1 - 99.8	67.9 (19/28)	47.6 - 84.1	2.9	0.098
BI-RADS R4 >3	100 (15/15)	78.2 - 100	64.3 (18/28)	59.5 - 98.3	2.8	<0.01
Tree R1 >4	100 (15/15)	78.2 - 100	60.7 (17/28)	40.6 - 78.5	2.55	<0.01
Tree R2 >4	100 (15/15)	78.2 - 100	57.1 (16/28)	37.2 – 75.5	2.33	<0.01
Tree R3 >4	100 (15/15)	78.2 - 100	60.7 (17/28)	40.6 - 78.5	2.55	<0.01
Tree R4 >4	100 (15/15)	78.2 - 100	67.9 (19/28)	47.6 - 84.1	3.11	<0.01

RADS, *Tree* provides specific guidance about what certain combinations of lesion features indicate with regard to potential malignancy. This simplifies and structures the process of lesion interpretation. Our results demonstrate a reduction of inter-observer variability related to reader experience. The *Tree* uses a small number of diagnostically relevant criteria in a simple flow-chart to lead the reader toward a definite diagnosis. The Oxford Centre for Evidence-Based Medicine requires diagnostic scoring systems such as Tree to be validated in an independent study in order to prove its diagnostic value [14]. Our study follows this recommendation to validate *Tree* in an independent study, taking place in another department using different MRI technology and readers.

For masses, the MRI BI-RADS terminology is effective in predicting malignancy and has a good reproducibility for the final category assignment [23–26]. Similar results could not be confirmed for non-mass lesions: it has been reported that BI-RADS descriptors fail to diagnose correctly non-mass lesions and that several morphologic and dynamic features show an overlap of diagnostic information [9, 23, 24, 26]. Consequently, there is a wide variability among radiologists in choosing the best BI-RADS lesion description, especially when reporting non-mass lesions. In addition, the diagnostic performance of MRI BI-RADS reading is affected by reader experience, as demonstrated by our results and a prior study. This prior study reported on poorest diagnostic outcomes in interpretation of MRI if less experienced readers assessed non-mass lesions [25]. Our findings show similar results for *Tree* reading in mass and non-mass lesions. Of note, all readers achieved 100 % sensitivity in non-mass lesions while specificity stayed similar or improved with *Tree*. This improvement did not show statistical significance due to the low number of cases, but, similar to mass lesions, the improved performance was strongest in the inexperienced reader. While the original *Tree* article did not investigate this issue [15], we demonstrated that *Tree* is applicable not only to

masses, but also to non-mass lesions. Moreover, inter-reader agreement was improved by using *Tree*. As a result, *Tree* can be applied without restrictions in the diagnostic setting.

Some limitations of our research merit acknowledgment. One would expect a certain bias toward higher inter-reader agreement by the mono-centric design of this study. However, all four readers were trained in four different institutions and underwent only a short training session, as described above. Therefore, our results clearly demonstrate the high reproducibility of *Tree*, which was superior to the BI-RADS reading approach. It is not our intent to replace the BI-RADS lexicon. On the contrary: *Tree* is complementary to BI-RADS, as it provides empirically validated guidance where no specific recommendations are contained in BI-RADS. We validated the previously proposed *Tree* classification algorithm. Both the initial study and this validation study focused on histologically verified lesions. Although we thus confirmed the applicability of *Tree* and demonstrated its use in mass and non-mass lesions alike, our results may not directly apply to a general population. Here, because of the large number of negative MRI cases not referred for biopsy, specificity is likely to be higher [27]. Our study was performed considering MRI features only, and did not integrate patient characteristics. Such an approach yielded higher diagnostic accuracy in a prior study on non-mass lesions [28]. Of note, the *Tree* system allows integration of further diagnostic data, as discussed in [15]. In addition to clinical and conventional findings, quantitative information, e.g., from DWI, may be integrated [29, 30]. However, this approach was beyond the aim of our study and has not been validated as yet.

In conclusion, the *Tree* scoring system improves inter-reader agreement and achieves a diagnostic performance similar to that of BI-RADS. Less experienced radiologists, in particular, benefit from *Tree*.

Acknowledgments The scientific guarantor of this publication is Pascal A. T. Baltzer. The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article. The authors state that no funding was received for this work. One of the authors has significant statistical expertise. Institutional Review Board approval was obtained. Written informed consent was waived by the Institutional Review Board. Methodology: retrospective, diagnostic or prognostic study, performed at one institution.

References

1. Warner E, Messersmith H, Causer P et al (2008) Systematic review: using magnetic resonance imaging to screen women at high risk for breast cancer. Ann Intern Med 148:671–679
2. Riedl CC, Luft N, Bernhart C et al (2015) Triple-modality screening trial for familial breast cancer underlines the importance of magnetic resonance imaging and questions the role of mammography and ultrasound regardless of patient mutation status, age, and breast density. J Clin Oncol Off J Am Soc Clin Oncol. doi:10.1200/JCO.2014.56.8626
3. Houssami N, Ciatto S, Macaskill P et al (2008) Accuracy and surgical impact of magnetic resonance imaging in breast cancer staging: systematic review and meta-analysis in detection of multifocal and multicentric cancer. J Clin Oncol Off J Am Soc Clin Oncol 26:3248–3258
4. Kuhl C (2007) The current status of breast MR imaging. Part I. Choice of technique, image interpretation, diagnostic accuracy, and transfer to clinical practice. Radiology 244:356–378
5. Sardanelli F, Boetes C, Borisch B et al (2010) Magnetic resonance imaging of the breast: recommendations from the EUSOMA working group. Eur J Cancer Oxf Engl 46:1296–1316
6. Mann RM, Balleyguier C, Baltzer PA, European Society of Breast Imaging (EUSOBI), with language review by Europa Donna–The European Breast Cancer Coalition et al (2015) Breast MRI: EUSOBI recommendations for women's information. Eur Radiol. doi:10.1007/s00330-015-3807-z
7. The American College of Radiology (ACR) (2013) Breast Imaging Reporting and Data System Atlas (BI-RADS® Atlas), Reston, VA
8. Pinker K, Bogner W, Baltzer P et al (2014) Improved differentiation of benign and malignant breast tumors with multiparametric 18Fluorodeoxyglucose positron emission tomography magnetic resonance imaging: a feasibility study. Clin Cancer Res. doi:10.1158/1078-0432.CCR-13-2810
9. Benndorf M, Baltzer PAT, Kaiser WA (2011) Assessing the degree of collinearity among the lesion features of the MRI BI-RADS lexicon. Eur J Radiol 80:e322–e324
10. Ikeda DM, Hylton NM, Kinkel K et al (2001) Development, standardization, and testing of a lexicon for reporting contrast-enhanced breast magnetic resonance imaging studies. J Magn Reson Imaging 13:889–895
11. Kim SJ, Morris EA, Liberman L et al (2001) Observer variability and applicability of BI-RADS terminology for breast MR imaging: invasive carcinomas as focal masses. AJR Am J Roentgenol 177:551–557
12. Kinkel K, Helbich TH, Esserman LJ et al (2000) Dynamic high-spatial-resolution MR imaging of suspicious breast lesions: diagnostic criteria and interobserver variability. AJR Am J Roentgenol 175:35–43
13. Stoutjesdijk MJ, Fütterer JJ, Boetes C et al (2005) Variability in the description of morphologic and contrast enhancement characteristics of breast lesions on magnetic resonance imaging. Invest Radiol 40:355–362
14. OCEBM Levels of Evidence Working Group. "The Oxford Levels of Evidence 2". Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653
15. Baltzer PAT, Dietzel M, Kaiser WA (2013) A simple and robust classification tree for differentiation between benign and malignant lesions in MR-mammography. Eur Radiol 23:2051–2060
16. Baum F, Fischer U, Vosshenrich R, Grabbe E (2002) Classification of hypervascularized lesions in CE MR imaging of the breast. Eur Radiol 12:1087–1092
17. Nunes LW, Schnall MD, Orel SG et al (1997) Breast MR imaging: interpretation model. Radiology 202:833–841
18. Tozaki M, Igarashi T, Matsushima S, Fukuda K (2005) High-spatial-resolution MR imaging of focal breast masses: interpretation model based on kinetic and morphological parameters. Radiat Med 23:43–50
19. Tozaki M, Fukuda K (2006) High-spatial-resolution MRI of non-masslike breast lesions: interpretation model based on BI-RADS MRI descriptors. AJR Am J Roentgenol 187:330–337
20. Demartini WB, Kurland BF, Gutierrez RL et al (2011) Probability of malignancy for lesions detected on breast MRI: a predictive model incorporating BI-RADS imaging features and patient characteristics. Eur Radiol 21:1609–1617
21. Pinker K, Bogner W, Baltzer P et al (2014) Improved diagnostic accuracy with multiparametric magnetic resonance imaging of the

breast using dynamic contrast-enhanced magnetic resonance imaging, diffusion-weighted imaging, and 3-dimensional proton magnetic resonance spectroscopic imaging. Invest Radiol. doi:10.1097/RLI.0000000000000029

22. Perry N, Broeders M, de Wolf C et al (2008) European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. Ann Oncol 19:614–622
23. Gutierrez RL, DeMartini WB, Eby PR et al (2009) BI-RADS lesion characteristics predict likelihood of malignancy in breast MRI for masses but not for nonmasslike enhancement. AJR Am J Roentgenol 193:994–1000
24. Baltzer PAT, Benndorf M, Dietzel M et al (2010) False-positive findings at contrast-enhanced breast MRI: a BI-RADS descriptor study. AJR Am J Roentgenol 194:1658–1663
25. Baltzer PAT, Kaiser WA and Dietzel M (2015) Lesion type and reader experience affect the diagnostic accuracy of breast MRI: a multiple reader ROC study. Eur J Radiol 84(1):86–91. doi:10.1016/j.ejrad.2014.10.023.
26. Jansen SA, Shimauchi A, Zak L et al (2011) The diverse pathology and kinetics of mass, nonmass, and focus enhancement on MR imaging of the breast. J Magn Reson Imaging 33:1382–1389
27. Dietzel M, Baltzer PAT, Schön K, Kaiser WA (2012) MR-mammography: high sensitivity but low specificity? New thoughts and fresh data on an old mantra. Eur J Radiol 81:S30–S32
28. Thomassin-Naggara I, Trop I, Chopier J et al (2011) Nonmasslike enhancement at breast MR imaging: the added value of mammography and US for lesion categorization. Radiology 261:69–79
29. Pinker K, Bickel H, Helbich TH et al (2013) Combined contrast-enhanced magnetic resonance and diffusion-weighted imaging reading adapted to the "Breast Imaging Reporting and Data System" for multiparametric 3-T imaging of breast lesions. Eur Radiol 23:1791–1802
30. Baltzer A, Dietzel M, Kaiser CG, Baltzer PA (2015) Combined reading of contrast enhanced and diffusion weighted magnetic resonance imaging by using a simple sum score. Eur Radiol. doi:10.1007/s00330-015-3886-x