Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors

Ashirbani Saha, and Michael R. Harowicz

Department of Radiology, Duke University School of Medicine, 2424 Erwin Road, Suite 302, Durham, NC 27705, USA

Maciej A. Mazurowski^{a)}

Department of Radiology, Duke University School of Medicine, 2424 Erwin Road, Suite 302, Durham, NC 27705, USA

Department of Electrical and Computer Engineering, Duke University, Box 90291, Durham, NC 27708, USA

Duke University Medical Physics Program, DUMC 2729, 2424 Erwin Road, Suite 101, Durham, NC 27705, USA

(Received 26 July 2017; revised 1 March 2018; accepted for publication 4 April 2018; published 11 May 2018)

Purpose: To review features used in MRI radiomics of breast cancer and study the inter-reader stability of the features.

Methods: We implemented 529 algorithmic features that can be extracted from tumor and fibroglandular tissue (FGT) in breast MRIs. The features were identified based on a review of the existing literature with consideration of their usage, prognostic ability, and uniqueness. The set was then extended so that it comprehensively describes breast cancer imaging characteristics. The features were classified into 10 groups based on the type of data used to extract them and the type of calculation being performed. For the assessment of inter-reader variability, four fellowship-trained readers annotated tumors on preoperative dynamic contrast-enhanced MRIs for 50 breast cancer patients. Based on the annotations, an algorithm automatically segmented the image and extracted all features resulting in one set of features for each reader. For a given feature, the inter-reader stability was defined as the intraclass correlation coefficient (ICC) computed using the feature values obtained through all readers for all cases.

Results: The average inter-reader stability for all features was 0.8474 (95% CI: 0.8068–0.8858). The mean inter-reader stability was lower for tumor-based features (0.6348, 95% CI: 0.5391–0.7257) than FGT-based features (0.9984, 95% CI: 0.9970–0.9992). The feature group with the highest inter-reader stability quantifies breast and FGT volume. The feature group with the lowest inter-reader stability quantifies variations in tumor enhancement.

Conclusions: Breast MRI radiomics features widely vary in terms of their stability in the presence of inter-reader variability. Appropriate measures need to be taken for reducing this variability in tumor-based radiomics. © 2018 American Association of Physicists in Medicine [https://doi.org/10.1002/mp.12925]

Key words: breast cancer MRI, image segmentation, imaging features, inter-reader variability, intraclass correlation coefficient, radiogenomics, radiomics

1. INTRODUCTION

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is often acquired for breast cancer patients as it offers additional information to guide treatment.^{1,2} In comparison to mammography and ultrasound, it also expresses superior sensitivity for detection of additional or residual disease.^{1–4} Recently, computer algorithms have been applied to automatically extract large numbers of quantitative features that thoroughly describe various characteristics of tumors and their surroundings. These features have been applied to the determination of patient diagnosis,^{5,6} prognosis of outcomes,^{7–10} and for their association with genomics.^{11–14} These analyses are currently referred to as radiomics¹⁵ and radiogenomics.¹⁶

For imaging features related to breast cancer MRI, recent years have witnessed an ever-growing literature on feature extraction methodologies, potential usage of features in clinical context, and on uniqueness of the information offered by

a particular feature. These features have been typically developed using different datasets, across different institutions with varying image preprocessing techniques. A few examples of this diversity are Refs. [9,10,12,17–19]. Also, different features may capture similar information despite having technically different definitions for extraction. Most research groups use a relatively small subset of all features available in the literature. This necessitates the evaluation of the interrelation of these features in a uniform experimental setting.

In breast cancer MRI radiomics, the feature extraction process typically requires an annotation of the tumor by a radiologist^{9,20–24} making the tumor segmentation process either completely manual or semiautomatic. It is well established that radiologists vary in their annotation of medical images and particularly in annotation of tumors in breast MRIs.^{25,26} Difference in reader's annotation can introduce differences in the tumor masks (a binary image specifying the location and extent of the tumor), which would cause radiologist-specific tumor and fibroglandular tissue (FGT) masks to be

nonidentical. This will cause features (extracted using tumor and/or FGT) to be different for different readers. Thus inter-reader variability affects the values of different semiautomatic features that are extracted from tumor and FGT of the breast. What is not known is the extent to which different features are affected by the inter-reader variability. This issue is of utmost importance, since it could significantly affect the generalizability of the presented results and predictive models between institutions and in turn could be detrimental to the implementation of radiomics models in clinics.

While inter-reader variability (also known as interobserver variability) has been studied in the general context of medical imaging, little work is available on its impact in the rapidly growing field of radiomics. To our knowledge, this is the first study of impact of interobserver variability in breast MRI radiomics. There are only two prior studies^{25,27} showing the impact of interobserver variability on one feature implemented in their respective groups. In this study, we address this issue thoroughly by evaluating a very large set of radiomics features (529) developed by different groups working on radiomics including ours. We categorized the features based on the information used to calculate them as well as the type of calculation performed to provide an additional clarity to the field.

2. MATERIALS AND METHODS

2.A. Patient population

In this institutional review board-approved study, we randomly selected a subset of 50 invasive breast cancer patients that had preoperative MRI available from September 2007 to June 2009, exactly one biopsy proven cancer, and a supporting pathology report having the corresponding clock position of the biopsied tumor. The selected patients did not have history of breast cancer or elective breast surgery prior to MRI, and were not undergoing breast cancer treatment at the time of the MRI. The sample size of 50 is chosen such that it allows us to draw statistical conclusion and save the radiologists' time effort.

2.B. Imaging data and image annotation

The breast MRIs were axial and were acquired by 1.5 T or 3.0 T scanners in the prone position. For all the cases included in our experiments, the following MRI sequences were available: a fat-saturated gradient echo T1-weighted precontrast sequence, typically with four postcontrast T1-weighted sequences acquired after the IV administration of contrast agent (using a weight-based protocol of 0.2 mL/kg), and a nonfat-saturated T1-weighted sequence. For the fat-saturated MR sequences, the repetition time varied between 3.9 and 6.4 ms, echo time varied between 1.4 and 2.5 ms, and slice thickness was 1.1 or 2 mm. The distribution of patients for different scanners and contrast agents is shown in Table I.

The images were annotated by four fellowship-trained breast imagers (6 months–22 yr of postfellowship

TABLE I. Distribution of patients—scanner wise and contrast agent wise.

Characteristic	Technical details	Manufacturer details	Patient count
Magnetic field strength	1.5 T	Signa HDxt, GE Healthcare, Little Chalfont, UK	1
		Signa HDx, GE Healthcare, Little Chalfont, UK	11
		Avanto, Siemens, Munich, Germany	7
	3.0 T	Signa HDx, GE Healthcare, Little Chalfont, UK	4
		MAGNETOM Trio, Siemens, Munich, Germany	27
Contrast agent	Gadopentetate dimeglumine	Magnevist, Bayer Healthcare, Berlin, Germany	44
	Gadobenate dimeglumine	MultiHance, Bracco, Milan, Italy	2
	Not available	—	4

experience). Using a graphical user interface developed in our laboratory, the following data was displayed to the reader: (a) the precontrast sequence, (b) the first postcontrast sequence, (c) the subtracted sequence (computed by subtracting the precontrast from the first postcontrast sequence), and (d) the clock face location of the biopsied tumor. For all cases, the readers annotated the tumor using a bounding box on a slice, and indicated the starting and ending slices for the extent of the box. Each reader annotated all cases in a single session in a location which had an ambience mimicking a clinical reading room.

2.C. Image segmentation

Using a reader's annotation, we applied a fuzzy C-means automatic segmentation to obtain the tumor mask. For each patient and individual reader combination, two different FGT masks were extracted based on two different sequences because some techniques in the literature use a T1-nonfat-saturated sequence while others use a fat-saturated sequence.²⁸ The first FGT mask was automatically extracted from the N4-corrected²⁹ T1-nonfat-saturated (T1-NFS) images using the following steps: (a) segmentation of breast from the chest cavity to obtain breast mask,⁹ (b) segmentation of FGT using fuzzy C-means clustering,³⁰ (c) registration of the extracted mask to the DCE-MRI sequence, and (d) removal of any part of the segmented FGT that overlapped with segmented tumor to extract the final FGT (excluding the tumor) only. The second mask was extracted from the first T1-fat-saturated postcontrast (PostCon) sequence by (a) registration of the breast mask extracted from T1-NFS to PostCon (b) FGT segmentation, and (c) removal of any part of the segmented FGT that overlapped with the corresponding segmented tumor to extract the final FGT (excluding the tumor) only. Thus for each patient and reader combination, we have a unique set of masks for DCE-MRI: (a) breast mask (registered), (b) tumor mask, (c) FGT mask from T1-NFS, and (d)

FGT mask from PostCon. Note that, all of these preprocessing steps are automatic, and success of each step is required for the final extraction of features. We excluded one case from feature extraction, as it failed in one of these steps.

2.D. Selection and organization of imaging features

For this study and subsequent evaluation, a set of breast cancer radiomics features needed to be selected as there are a vast number of features existing in literature. We categorized the features by exploring their sources and the type of information they represent. For each category, we selected features from the existing literature, made amendments to their definitions to suit our study (when required), and added features to a category (when necessary) to leverage the use of the avalanche of information carried in breast MRI. The process of selection, categorization, and appending new features was iterative in nature.

Radiomic and radiogenomic feature extraction requires the use of breast masks, FGT masks, and tumor masks. There are some features, which require the masks only and not the intensity values of the MRI sequences for feature extraction. Hence, we divide the set of features based on the sequence dependence after mask extraction. Breast and FGT volume features (breast volume, FGT volume, and FGT density^{31,32}) along with some general features of tumor size and morphology^{11,33–35} belong to this category and are not affected by the intensity values of the sequences.

All other features use postcontrast sequences or a subset of postcontrast sequences (along with the optional use of the pre-contrast sequence) and therefore are called the enhancement and kinetic features. In the literature, the tumor enhancement-related features have found significant importance due to their relationships to diagnosis, prognosis of outcomes, and genomic characteristics. These features include different types of characterization of tumor enhancement. Here, we have volumetric and mean enhancement features computed by grouping tumor voxels according to the enhancement time point²⁴ and voxel-wise enhancement values to form signal enhancement ratio (SER) maps, and peak enhancement (PE) maps.³⁶ Other features compute different measures of dispersions from these aforementioned groups and maps and fall in the distinct category of tumor enhancement variation. Tumor enhancement variation is also captured by the features that compute different properties of the tumor uptake curves.^{18,37,38} Interestingly, these features of tumor enhancement variation do not use the spatial relationship of the tumor voxels.

The features of tumor enhancement that use the spatial relationship between voxels are categorized in either of two groups: (a) tumor enhancement texture features or (b) tumor enhancement spatial heterogeneity features. The texture features¹⁹ are commonly used in tumor radiomics/radiogenomics to compute the spatial relationships between the enhancement patterns of the tumor. In our set of features, these texture features³⁹ are computed from the first postcontrast, PE, SER rate maps, and washing rate maps. Other texture features related to discrete Fourier transform (DFT),⁵

dynamic local binary patterns (DLBP), and dynamic histogram of oriented gradients (DHOG)⁴⁰ are also included. Furthermore, spatial relationships of tumor enhancement variation along the tumor margin are also included in this group.¹⁸ While the texture features that we included are indicative of spatial variations in enhancement in general, we have included four features of spatial heterogeneity of enhancement computed using global Moran's I⁴¹ and clustering enhancements.

In our study, we extracted features related to the proportion of enhanced FGT.⁴² These features convey information about background parenchymal enhancement (BPE). As several recent studies^{17,43–45} have indicated the potential effectiveness of FGT enhancement-based biomarkers in prognosis and outcomes, we have assigned substantial importance to the FGT-based features. Several FGT enhancement texture and FGT enhancement variation features are computed, using one of the FGT masks at a time. Some of these features are inspired by corresponding features that are computed using tumor masks only. Another set of features that combines both tumor and FGT enhancement dynamics^{12,13} was also included in our feature set. Finally, we had 529 features grouped into 10 categories. Figure 1 shows the groups and the number of features belonging to each of the groups.

The number of features based on FGT is higher because they were extracted from the two types of FGT masks. The complete list of all features that we extracted, along with the related references and modifications, can be found in the Appendix (Supplementary Material A) of our previous study.⁴⁶ Some of these features were based on the algorithms proposed by our group. A majority of the features extraction algorithms were proposed by other researchers. These features were reimplemented in-house in Matlab 2016b (The Mathworks, Natick, MA).

2.E. Analysis of inter-reader stability

For four readers' annotations, four sets of the selected features were obtained for comparison. In the first step, we analyzed each feature by defining a measure of its inter-reader stability. In the second step of our analysis, we quantified how the correlations (Pearson's linear correlation coefficient) of the feature groups changed due to the variability between readers. In the third step, we quantified how pairwise reader variability translated to the variability of the feature groups.

2.E.1. Measuring inter-reader stability of features

As studies in literature use intraclass correlation (ICC) for the evaluation of stability/reliability/reproducibility of features,^{47–50} we have evaluated the ICC(3,1) (consistency of the reader by two-way model using the *irr* package in R⁵¹) values for the each of the 529 features. Please note that inter-reader stability and inter-reader variability are inversely related. Higher inter-reader stability (computed in terms of ICC) of a feature implies that the effect of inter-reader variability on the feature is low.

FIG. 1. Flowchart demonstrating the organization of features and grouping used in the analysis.

2.E.2. Inter-relationships within the feature groups

We computed the Pearson's linear correlation coefficient (PLCC) between all pairs of 529 features corresponding to the same reader. Then according to the grouping, the average intergroup and intragroup PLCC values were computed. This was repeated for all readers to visualize the change in inter-relation of the features groups caused by the reader variability.

2.E.3. Effect of pairwise reader variability on feature stability

To evaluate how the pairwise inter-reader variability affected the feature stability, we calculated the average Dice similarity coefficients between each pair of readers' 3D annotation (bounding boxes) and the average ICC values for each feature group for a pair of readers.

3. RESULTS

The inter-reader stability for each feature is shown in Fig. 2. The bars in the figure indicate the value of ICC. Under each individual group, the features are organized in increasing order of inter-reader stability.

Please refer to the Supporting information of this paper for the stability values related to individual features. The effect of reader variability on different feature groups were assessed by the inter-reader stability values of the individual features included in the respective feature groups. We considered an inter-reader stability (ICC) value greater than 0.9 as highly stable features, similar to the work investigating inter-reader stability of diffusion weighted breast MRI.⁴⁸ However, the same work did not specify any lower limit of ICC for identifying features with low stability. Therefore,

based on the studies,^{52,53} features with stability values less than 0.7 are considered having low stability. Among the tumor features, we found a set of highly stable features related to different groups. Among the size and morphological features, tumor volume was highly stable. Several features of tumor enhancement related to grouping of the tumor based on washin maps, washout maps, peak enhancement, SER maps, and tumor uptake demonstrated high inter-reader stability. However, not every feature related to these maps or groups had high stability. Approximately, 10% of the tumor features are highly stable. Inter-reader stability values of less than 0.7 were obtained for about 54% of the tumor-based features. For the FGT-based features, the minimum inter-reader stability is 0.94, and therefore, all FGT-related features are highly stable.

An average inter-reader stability for all features of 0.8474 (95% CI: 0.8068–0.8858) was obtained for all features. The mean inter-reader stability was lower for tumor-based features (0.6348, 95% CI: 0.5391–0.7257) than FGT-based features (0.9984, 95% CI: 0.9970–0.9992). The average ICC values and 95% confidence intervals of 10 feature groups are shown in Table II. When computed for a single scanner (Signa HDx, GE Medical systems, $n = 38$), a slight improvement was obtained for inter-reader stability of tumor-based features (0.6456, 95% CI: 0.5498–0.7411) and all features (0.8528, 95% CI: 0.8127–0.8924), whereas minor decrease was obtained for FGT-based features (0.9979, 95% CI: 0.9959–0.9990).

The median and variation of inter-reader stability of each feature in each group is shown in Fig. 3. We find that in general, groups related to tumor-based features have lower inter-reader stability and more variation compared to groups related to FGT-based features. The feature group that combines features from tumor and FGT enhancements had a median inter-reader stability value in between that of the only

Fig. 2. Stability of each feature arranged in increasing order (from top to bottom) in each group.

TABLE II. Values of 95% confidence intervals for the average ICC values of each feature group.

Name of the group	Average ICC for the group	95% Confidence intervals for average groupwise ICC values
Breast and FGT volume features	0.9999	0.9998–1
Tumor size and morphology	0.6627	0.5242–0.7668
FGT enhancement	0.9994	0.9983–0.9999
Tumor enhancement	0.7678	0.6704–0.9158
Combining tumor and FGT enhancement	0.8839	0.7948–0.9445
FGT enhancement texture	0.9982	0.9970–0.9990
Tumor enhancement texture	0.6104	0.5002–0.7018
Tumor enhancement spatial heterogeneity	0.7819	0.6444–0.8702
FGT enhancement variation	0.9968	0.9908–0.9995
Tumor enhancement variation	0.5903	0.4646–0.7384

tumor- and only FGT-based feature groups with 50% highly stable and almost 6% with lower stability.

The groupwise inter-relation of features using PLCC is shown in Fig. 4. As we see in the within group, the inter-relation does not change significantly (maximum standard deviation 0.05) between readers. Within four readers, the intergroup PLCC varies most for tumor enhancement spatial heterogeneity features (standard deviation: 5.3%), followed by combining tumor and FGT enhancement (standard deviation: 4.4%), and tumor size and morphology (standard deviation: 2.7%). The average variation of PLCC for all pairs of

feature groups is less than 1.5%. All features groups (with the exception of tumor enhancement texture) had higher intragroup PLCC values than all intergroup PLCC values. Thus, features belonging to a group are more correlated with each other than with features from other groups for all readers.

The effect of the pairwise reader variability on the pairwise ICC is shown in Table III. The reader pair (reader 1 vs reader 3) that had highest similarity in the bounding boxes in terms of DICE similarity coefficient showed the highest mean ICC across all feature groups. All but one feature group had highest stability for this reader pair. This demonstrates that the inter-rater variability in annotation directly translates into differences in extracted features. Across all feature groups, lowest mean ICC was obtained for the reader pair (reader 2 vs reader 4) with the lowest similarity in annotation. Half of the tumor-related feature groups showed lowest stability for this reader pair and the mean ICC across all feature groups was also lowest for this pair.

4. DISCUSSION

In this study, we assembled a set of breast MRI radiomics and radiogenomics features from the existing literature, applied necessary modifications and extensions to build a set of features that captured a representative part of breast cancer characteristics, and categorized them. This analysis allowed us to identify aspects of computer vision-based analysis of breast MR images that have been more thoroughly covered such as tumor enhancement and those that were less explored such as features related to spatial heterogeneity in tumor

FIG. 3. Boxplot for the distribution of stability of features belonging different groups. [Color figure can be viewed at wileyonlinelibrary.com]

enhancement and features based on the combination of tumor and FGT enhancement. These two areas may benefit from additional analysis and development of new features.

Our feature selection procedure aimed at gathering features that were previously found effective for diagnosis, outcomes prediction, and radiogenomics. It is possible that we did not include features that were previously presented in the literature and did not demonstrate prognostic power in the dataset in which they were studied. Our goal was directed toward forming a comprehensive set of promising features for breast cancer MRI and to evaluate their stability in the presence of inter-reader variability. Evaluation of effectiveness of these features in clinical context on any independent dataset is an open problem. Once that evaluation is made, the inter-reader variability can help in highlighting features that have higher stability and higher applicability in clinical context in order to form a stable set of breast MRI radiomics and radiogenomics features.

We found that FGT-related features have significantly higher stability in the presence of inter-reader variability as opposed to tumor-based features. The main reason for this is that the FGT mask is not highly affected by the reader's annotation of the tumor. Typically, the volume of the FGT is significantly higher than that of the tumor. The proportion of change in FGT mask (which considers FGT by removing the tumor mask-related voxels) is much lower than that in tumor mask for different readers. Hence, the features related to FGT mask can endure changes due to inter-reader variability in tumor mask in a better manner. However, tumor features are more likely to be important clinically. Therefore, it is

important to reduce inter-reader variability in the tumor radiomics. One of the ways to do this is by streamlining the process of annotations with precise instructions such that the inter-reader variability can be minimized. In our earlier study,²⁵ we demonstrated that the agreement between readers is increased when automatic tumor segmentation is applied (change of average Dice coefficient from 0.60 to 0.77). Please note that the resulting moderate level of stability (0.77) can have different effect on stability of features. Those that are not strongly affected by tumor segmentation such as overall intensity features will not show high variability due to inter-reader variability. On the other hand, those that are strongly affected by a changing mask such as shape features of the tumor will show high variability.

In this study, we concluded that average inter-reader variability of semiautomatic imaging features for breast cancer MRI radiomics/radiogenomics is moderate. While this inter-reader variability can potentially be reduced with improved and precise annotation instructions, and by providing more contextual data to the readers, it remains an issue that should be carefully considered in radiomics and radiogenomics studies as it could affect generalizability of the results. These imaging features are extracted with the common goal of representing certain aspects of the disease under study. If a feature is significantly susceptible to subjective bias, its potential usage as a stable clinical biomarker becomes questionable. Therefore, assessing and reducing the effect of subjective bias for semiautomatic imaging biomarker is vital.

We carried our study on a set of 50 randomly selected patients. Increasing the number of patients might be

FIG. 4. Average inter- and intragroup Pearson's linear correlation coefficient (PLCC) of features for (a) reader 1 (b) reader 2 (c) reader 3, and (d) reader 4. Brighter shades indicate higher PLCC.

beneficial for a better analysis of inter-reader variability, although this dataset can be considered sufficient for the type of analysis conducted in this paper based on the following : (a) This set of patients has been used in our earlier analysis²⁵ and shown to have tumors of various characteristics (e.g., mass enhancements, nonmass enhancements, hematoma, multiple tumor masses in similar clock positions). (b) The tumor characteristics caused readers to vary in tumor annotation. As concluded in the study,⁵⁴ the quality of represented information should determine the sample size, the reader variability was captured through the set of 50 patients. (c) As an additional clarification, the study⁵⁵ developed a method for computing the sample size for inter-reader reliability studies using ICC. Based on their technique, using the thresholds (0.7 for low stability and 0.9 for high stability) chosen in our analysis for four readers, the optimal sample size that minimizes number of observations (number of readers times

number of subjects or cases) is 11.1. Therefore, the sample size of 50 used in our study is well above 11.1 to draw conclusions. Also, we had to make minor changes in the computation of some features (e.g., using three-dimensional tumor instead of central tumor slice) in order to make them applicable in our settings.

In the present literature related to inter-reader variability and MR extracted features, several studies^{56–59} have considered MR BI-RADS⁵⁶ based features and not radiomics features. Two studies^{25,27} have conducted inter-reader variability of the specific imaging features that they extracted as part of their analyses. A high inter-reader average correlation (PLCC) was obtained for the tumor- and FGT-related feature observed in Ref. [25] based on four readers. A moderate inter-reader stability (using ICC) was observed for the tumor-based feature extracted from T2-weighted MR sequences²⁷ based on two readers.

TABLE III. Pairwise reader variability versus stability of features.

	Reader 1 vs reader 2	Reader 1 vs reader 3	Reader 1 vs reader 4	Reader 2 vs reader 3	Reader 2 vs reader 4	Reader 3 vs reader 4
DICE coefficients for reader pair						
For bounding boxes	0.5925	0.7403	0.5915	0.6029	<u>0.5055</u>	0.6051
ICC for feature group						
Breast and FGT volume features	<u>0.9999</u>	1	<u>0.9999</u>	<u>0.9999</u>	1	1
Tumor size and morphology	0.8213	0.9017	0.7188	0.8207	<u>0.6952</u>	0.7083
FGT enhancement	0.9997	0.9999	<u>0.9995</u>	0.9997	<u>0.9997</u>	<u>0.9995</u>
Tumor enhancement	0.7991	0.9796	<u>0.7951</u>	0.8059	0.8489	0.7986
Combining tumor and FGT enhancement	0.9452	0.9872	0.9218	0.9353	<u>0.9004</u>	0.9311
FGT enhancement texture	0.9994	0.9994	<u>0.9986</u>	0.9994	0.9989	0.9991
Tumor enhancement texture	0.7787	0.8119	<u>0.6101</u>	0.8098	0.6360	0.6779
Tumor enhancement spatial heterogeneity	0.8472	0.9421	0.7876	0.9541	<u>0.7844</u>	0.8100
FGT enhancement variation	0.9995	0.9998	<u>0.9970</u>	0.9996	0.9973	<u>0.9970</u>
Tumor enhancement variation	0.6242	0.9197	0.6628	0.6417	<u>0.5494</u>	0.6303
Mean ICC	0.8814	0.9541	0.8491	0.8966	<u>0.8410</u>	0.8552

Bold indicates the highest value in the row and underline indicates the lowest value in the row.

The intention of our study was to determine the impact of inter-reader variability on extraction of the radiomic features in general. Therefore, we did not consider specific uses of the radiomics features such as prediction of malignancy, distinguishing different types of tumors (e.g., DCIS vs invasive cancer), prognosis of outcomes (e.g., distant recurrence-free survival or response to neoadjuvant therapy), or a determining genomic properties of the tumors (e.g., molecular subtype or the Oncotype DX recurrence score). It is quite likely that different tasks will find different sets of features more or less useful and future studies should carefully investigate the compounding effects of inter-reader variability and predictive values of the features for specific tasks.

Our study had limitations. One limitation in this study is that the robustness of the features is considered in terms of the inter-reader stability values. The individual predictive abilities of the features are not considered in this work and can be done as a future study. However, a feature with high stability value implies that its predictive ability is less affected by inter-reader variability. While we have tried to build a comprehensive set of features describing breast cancer MRI, this set can always be extended and more features can be included as the breast MRI radiomics is constantly expanding.^{60–62} To be comprehensive, we included features of the breast as whole, FGT, and tumors, and considered features that quantify volume, shape, enhancement, and heterogeneity. We included both, features that we proposed in our laboratory and features proposed by other groups. However, not all features used in other studies were incorporated in our work. The primary reason for not including a feature was that a similar feature already existed in our set to limit redundancy of our feature set.

This study focuses on a single common combination of the reader annotation format and the segmentation algorithm. We believe that this common combination is an appropriate choice for this study and allows for better understanding of

the relationship between reader variability and radiomics feature stability. However, further studies in which the data is collected from radiologists in a different format (e.g., individual points) and different algorithms are applied (e.g., region growing) would be of benefit to understand how well the relationships found in this study generalize to other methods.

Finally, we calculated and listed inter-reader stability for all 529 features analyzed here (Supporting information). This study can serve as a guide for other researchers who might choose to include only those features in their study that demonstrate a high level of stability or interpret the results of their radiomics or radiogenomics studies in the context of feature stability.

5. CONCLUSIONS

As assessed by our analysis, inter-reader variability widely affects the stability of radiomics features from breast MRI. Reducing this variability in tumor-based radiomics by appropriate measures is necessary as these features are affected more compared to the FGT-based radiomics, and have found more importance in literature for diagnosis, genomic associations, and prognosis.

FUNDING

National Institutes of Health, 1R01EB021360(MAM). North Carolina Biotechnology Center, 2016-BIG-6520(MAM).

CONFLICTS OF INTEREST

The authors have no conflicts of interests to disclose.

^{a)} Author to whom correspondence should be addressed. Electronic mail: maciej.mazurowski@duke.edu; Telephone: +919 684 1466.

REFERENCES

- Knuttel FM, Menezes GLG, van den Bosch MAAJ, Gilhuijs KGA, Peters NHGM. Current clinical indications for magnetic resonance imaging of the breast. *J Surg Oncol*. 2014;110:26–31.
- Brasic N, Wisner DJ, Joe BN. Breast MR imaging for extent of disease assessment in patients with newly diagnosed breast cancer. *Magn Reson Imaging Clin N Am*. 2013;21:519–532.
- Shin HJ, Kim HH, Ahn JH, et al. Comparison of mammography, sonography, MRI and clinical examination in patients with locally advanced or inflammatory breast cancer who underwent neoadjuvant chemotherapy. *Br J Radiol*. 2011;61:612–620.
- Belli P, Costantini M, Malaspina C, Magistrelli A, LaTorre G, Bonomo L. MRI accuracy in residual disease evaluation in breast cancer patients treated with neoadjuvant chemotherapy. *Clin Radiol*. 2006;61:946–953.
- Zheng Y, Englander S, Baloch S, et al. STEP: spatiotemporal enhancement pattern for MR-based breast tumor diagnosis. *Med Phys*. 2009;36:3192–3204.
- Yang Q, Li L, Zhang J, Shao G, Zheng B. A new quantitative image analysis method for improving breast cancer diagnosis using DCE-MRI examinations. *Med Phys*. 2015;42:103–109.
- Partridge SC, Gibbs JE, Lu Y, et al. MRI measurements of breast tumor volume predict response to neoadjuvant chemotherapy and recurrence-free survival. *Am J Roentgenol*. 2005;184:1774–1781.
- Park VY, Kim E-K, Kim MJ, Yoon JH, Moon HJ. Breast parenchymal signal enhancement ratio at preoperative magnetic resonance imaging: association with early recurrence in triple-negative breast cancer patients. *Acta Radiol*. 2015;57:802–808.
- Mazurowski MA, Grimm LJ, Zhang J, et al. Recurrence-free survival in breast cancer is associated with MRI tumor enhancement dynamics quantified using computer algorithms. *Eur J Radiol*. 2015;84:2117–2122.
- Mahroogh M, Ashraf AB, Daye D, et al. Pharmacokinetic tumor heterogeneity as a prognostic biomarker for classifying breast cancer recurrence risk. *IEEE Trans Biomed Eng*. 2015;62:1585–1594.
- Sutton EJ, Oh JH, Dashevsky BZ, et al. Breast cancer subtype intertumor heterogeneity: MRI-based features predict results of a genomic assay. *J Magn Reson Imaging*. 2015;42:1398–1406.
- Mazurowski MA, Zhang J, Grimm LJ, Yoon SC, Silber JJ. Radiogenomic analysis of breast cancer: luminal B molecular subtype is associated with enhancement dynamics at MR imaging. *Radiology*. 2014;273:365–372.
- Grimm LJ, Zhang J, Mazurowski MA. Computational approach to radiogenomics of breast cancer: luminal A and luminal B molecular subtypes are associated with imaging features on routine breast MRI extracted using computer vision algorithms. *J Magn Reson Imaging*. 2015;42:902–907.
- Agner SC, Rosen MA, Englander S, et al. Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced MR images: a feasibility study. *Radiology*. 2014;272:91–99.
- Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48:441–446.
- Mazurowski MA. Radiogenomics: what it is and why it is important. *J Am Coll Radiol*. 2015;12:862–866.
- Wang J, Kato F, Oyama-Manabe N, et al. Identifying triple-negative breast cancer using background parenchymal enhancement heterogeneity on dynamic contrast-enhanced MRI: a pilot radiomics study. *PLoS ONE*. 2015;10:e0143308.
- Gilhuijs KGA, Giger ML, Bick U. Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging. *Med Phys*. 1998;25:1647–1654.
- Bhooshan N, Giger ML, Jansen SA, Li H, Lan L, Newstead GM. Cancerous breast lesions on dynamic contrast-enhanced MR images: computerized characterization for image-based prognostic markers. *Radiology*. 2010;254:680–690.
- Chen W, Giger ML, Bick U. A fuzzy C-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR Images I. *Acad Radiol*. 2006;13:63–72.
- Nie K, Chen J-H, Yu HJ, Chu Y, Nalcioğlu O, Su M-Y. Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI. *Acad Radiol*. 2008;15:1513–1525.
- Guo W, Li H, Zhu Y, et al. Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data. *J Med Imaging*. 2015;2:041007–041007.
- Cho N, Im S-A, Park I-A, et al. Breast cancer: early prediction of response to neoadjuvant chemotherapy using parametric response maps for MR imaging. *Radiology*. 2014;272:385–396.
- Ashraf AB, Daye D, Gavenonis S, et al. Identification of intrinsic imaging phenotypes for breast cancer tumors: preliminary associations with gene expression profiles. *Radiology*. 2014;272:374–384.
- Saha A, Grimm LJ, Harowicz M, et al. Interobserver variability in identification of breast tumors in MRI and its implications for prognostic biomarkers and radiogenomics. *Med Phys*. 2016;43:4558–4564.
- Beresford MJ, Padhani AR, Taylor NJ, et al. Inter- and intraobserver variability in the evaluation of dynamic breast cancer MRI. *J Magn Reson Imaging*. 2006;24:1316–1325.
- Henderson S, Purdie C, Michie C, et al. Interim heterogeneity changes measured using entropy texture features on T2-weighted MRI at 3.0 T are associated with pathological response to neoadjuvant chemotherapy in primary breast cancer. *Eur Radiol*. 2017;27:4602–4611.
- Chang DHE, Chen J-H, Lin M, et al. Comparison of breast density measured on MR images acquired using fat-suppressed versus nonfat-suppressed sequences. *Med Phys*. 2011;38:5961–5968.
- Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29:1310–1320.
- Bezdek JC. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA: Kluwer Academic Publishers; 1981.
- Klifa C, Carballido-Gamio J, Wilmes L, et al. Magnetic resonance imaging for secondary assessment of breast density in a high-risk cohort. *Magn Reson Imaging*. 2010;28:8–15.
- Faermann R, Sperber F, Schneebaum S, Barsuk D. Tumor-to-breast volume ratio as measured on MRI: a possible predictor of breast-conserving surgery versus mastectomy. *Isr Med Assoc J*. 2014;16:101–105.
- Giger ML, Vyborny CJ, Schmidt RA. Computerized characterization of mammographic masses: analysis of spiculation. *Cancer Lett*. 1994;77:201–211.
- Georgiou H, Mavroforakis M, Dimitropoulos N, Cavouras D, Theodoridis S. Multi-scaled morphological features for the characterization of mammographic masses using statistical classification schemes. *Artif Intell Med*. 2007;41:39–55.
- Czarnek N, Clark K, Peters KB, Mazurowski MA. Algorithmic three-dimensional analysis of tumor shape in MRI improves prognosis of survival in glioblastoma: a multi-institutional study. *J Neurooncol*. 2017;132:55–62.
- Arasu VA, Chen RCY, Newitt DN, et al. Can signal enhancement ratio (SER) reduce the number of recommended biopsies without affecting cancer yield in occult MRI-detected lesions? *Acad Radiol*. 2011;18:716–721.
- Chen W, Giger ML, Bick U, Newstead GM. Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI. *Med Phys*. 2006;33:2878–2887.
- Chen W, Giger ML, Lan L, Bick U. Computerized interpretation of breast MRI: investigation of enhancement-variance dynamics. *Med Phys*. 2004;31:1076–1082.
- Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. *IEEE Trans Syst Man Cybern*. 1973;3:610–621.
- Wan T, Bloch BN, Plecha D, et al. A radio-genomics approach for identifying high risk estrogen receptor-positive breast cancers on DCE-MRI: preliminary results in predicting OncotypeDX risk scores. *Sci Rep*. 2016;6:21394.
- Song C, Smith M, Huang Y, Jeraj R, Fain S. Heterogeneity of vascular permeability in breast lesions with dynamic contrast enhanced MRI. Paper presented at: 17th International Symposium for Magnetic Resonance in Medicine; 2009.
- Wu S, Weinstein SP, DeLeo MJ, et al. Quantitative assessment of background parenchymal enhancement in breast MRI predicts response to risk-reducing salpingo-oophorectomy: preliminary evaluation in a cohort of BRCA1/2 mutation carriers. *Breast Cancer Res*. 2015;17:1.
- King V, Brooks JD, Bernstein JL, Reiner AS, Pike MC, Morris EA. Background parenchymal enhancement at breast MR imaging and breast cancer risk. *Radiology*. 2011;260:50–60.
- Ahn HS, Kim SM, Jang M, Yun BL. Quantitative analysis of breast parenchymal background enhancement (BPE) on magnetic resonance

- (MR) imaging: association with mammographic breast density and aggressiveness of the primary cancer in postmenopausal women. *J Clin Oncol*. 2013;31:38–38.
45. Uematsu T, Kasami M, Watanabe J. Background enhancement of mammary glandular tissue on breast dynamic MRI: imaging features and effect on assessment of breast cancer extent. *Breast Cancer*. 2012;19:259–265.
 46. Saha A, Yu X, Sahoo D, Mazurowski MA. Effects of MRI scanner parameters on breast cancer radiomics. *Expert Syst Appl*. 2017;87:384–391.
 47. Leijenaar RTH, Carvalho S, Velazquez ER, et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol*. 2013;52:1391–1397.
 48. Spick C, Bickel H, Pinker K, et al. Diffusion-weighted MRI of breast lesions: a prospective clinical investigation of the quantitative imaging biomarker characteristics of reproducibility, repeatability, and diagnostic accuracy. *NMR Biomed*. 2016;29:1445–1453.
 49. Bogowicz M, Riesterer O, Bundschuh RA, et al. Stability of radiomic features in CT perfusion maps. *Phys Med Biol*. 2016;61:8736–8749.
 50. Parmar C, Rios Velazquez E, Leijenaar R, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS ONE*. 2014;9:e102107.
 51. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013.
 52. McClellan HL, Sakalidis VS, Hepworth AR, Hartmann PE, Geddes DT. Validation of nipple diameter and tongue movement measurements with B-mode ultrasound during breastfeeding. *Ultrasound Med Biol*. 2010;36:1797–1807.
 53. Morris C, Kurinczuk JJ, Fitzpatrick R, Rosenbaum PL. Reliability of the manual ability classification system for children with cerebral palsy. *Dev Med Child Neurol*. 2006;48:950–953.
 54. Sandelowski M. Sample size in qualitative research. *Res Nurs Health*. 1995;18:179–183.
 55. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med*. 1998;17:101–110.
 56. Ikeda DM, Hylton NM, Kinkel K, et al. Development, standardization, and testing of a lexicon for reporting contrast-enhanced breast magnetic resonance imaging studies. *J Magn Reson Imaging*. 2001;13:889–895.
 57. Grimm LJ, Zhang J, Baker JA, Soo MS, Johnson KS, Mazurowski MA. Relationships between MRI breast imaging-reporting and data system (BI-RADS) lexicon descriptors and breast cancer molecular subtypes: internal enhancement is associated with luminal B subtype. *Breast J*. 2017;23:579–582.
 58. Wengert GJ, Helbich TH, Woitek R, et al. Inter- and intra-observer agreement of BI-RADS-based subjective visual estimation of amount of fibroglandular breast tissue with magnetic resonance imaging: comparison to automated quantitative assessment. *Eur Radiol*. 2016;26:3917–3922.
 59. El Khoury M, Lalonde L, David J, Labelle M, Mesurolle B, Trop I. Breast imaging reporting and data system (BI-RADS) lexicon for breast MRI: interobserver variability in the description and assignment of BI-RADS category. *Eur J Radiol*. 2015;84:71–76.
 60. Aghaei F, Tan M, Hollingsworth AB, Qian W, Liu H, Zheng B. Computer-aided breast MR image feature analysis for prediction of tumor response to chemotherapy. *Med Phys*. 2015;42:6520–6528.
 61. Fan M, Wu G, Cheng H, Zhang J, Shao G, Li L. Radiomic analysis of DCE-MRI for prediction of response to neoadjuvant chemotherapy in breast cancer patients. *Eur J Radiol*. 2017;94:140–147.
 62. Wu J, Sun X, Wang J, et al. Identifying relations between imaging phenotypes and molecular subtypes of breast cancer: model discovery and external validation. *J Magn Reson Imaging*. 2017;46:1017–1027.

SUPPORTING INFORMATION

Additional supplemental material may be found online in the Supporting Information section at the end of the article.

Appendix S1. Supporting information for individual features.