

Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Breast MRI radiomics for the pretreatment prediction of response to neoadjuvant chemotherapy in node-positive breast cancer patients

Karen Drukker
Alexandra Edwards
Christopher Doyle
John Papaioannou
Kirti Kulkarni
Maryellen L. Giger

SPIE.

Karen Drukker, Alexandra Edwards, Christopher Doyle, John Papaioannou, Kirti Kulkarni, Maryellen L. Giger, "Breast MRI radiomics for the pretreatment prediction of response to neoadjuvant chemotherapy in node-positive breast cancer patients," *J. Med. Imag.* **6**(3), 034502 (2019), doi: 10.1117/1.JMI.6.3.034502.

Breast MRI radiomics for the pretreatment prediction of response to neoadjuvant chemotherapy in node-positive breast cancer patients

Karen Drukker,* Alexandra Edwards, Christopher Doyle, John Papaioannou, Kirti Kulkarni, and Maryellen L. Giger

University of Chicago, Department of Radiology, Chicago, Illinois, United States

Abstract. The purpose of this study was to evaluate breast MRI radiomics in predicting, prior to any treatment, the response to neoadjuvant chemotherapy (NAC) in patients with invasive lymph node (LN)-positive breast cancer for two tasks: (1) prediction of pathologic complete response and (2) prediction of post-NAC LN status. Our study included 158 patients, with 19 showing post-NAC complete pathologic response (pathologic TNM stage T0,N0,MX) and 139 showing incomplete response. Forty-two patients were post-NAC LN-negative, and 116 were post-NAC LN-positive. We further analyzed prediction of response by hormone receptor subtype of the primary cancer (77 hormone receptor-positive, 39 HER2-enriched, 38 triple negative, and 4 cancers with unknown receptor status). Only pre-NAC MRIs underwent computer analysis, initialized by an expert breast radiologist indicating index cancers and metastatic axillary sentinel LNs on DCE-MRI images. Forty-nine computer-extracted radiomics features were obtained, both for the primary cancers and for the metastatic sentinel LNs. Since the dataset contained MRIs acquired at 1.5 T and at 3.0 T, we eliminated features affected by magnet strength using the Mann–Whitney U-test with the null-hypothesis that 1.5 T and 3.0 T samples were selected from populations having the same distribution. Bootstrapping and ROC analysis were used to assess performance of individual features in the two classification tasks. Eighteen features appeared unaffected by magnet strength. Pre-NAC tumor features generally appeared uninformative in predicting response to therapy. In contrast, some pre-NAC LN features were able to predict response: two pre-NAC LN features were able to predict pathologic complete response (area under the ROC curve (AUC) up to 0.82 [0.70; 0.88]), and another two were able to predict post-NAC LN-status (AUC up to 0.72 [0.62; 0.77]), respectively. In the analysis by a hormone receptor subtype, several potentially useful features were identified for predicting response to therapy in the hormone receptor-positive and HER2-enriched cancers. © 2019 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.6.3.034502]

Keywords: radiomics; CAD; precision medicine; breast cancer; MRI.

Paper 19099RR received Apr. 11, 2019; accepted for publication Sep. 9, 2019; published online Sep. 30, 2019.

1 Introduction

There is a large variation in the clinical presentation of, and outcome of, breast cancer in women. It has been shown that in many instances biological biomarkers, i.e., features, of the primary tumor correlate with outcome. The availability of biomarkers that can be used to assess outcome as early and as accurately as possible is crucial to the development of successful targeted breast cancer therapies. Tumor response to preoperative chemotherapy correlates with outcome and could be a surrogate for evaluating the effect of chemotherapy on micrometastases and rate of recurrence. Methods to assess biological biomarkers for the prediction of outcome, however, may be invasive, expensive, not repeatable, or not widely available. Our hypothesis is that MR image-based features obtained through computer-extracted radiomics, an extension of computer-aided diagnosis, will prove useful as noninvasive biomarkers for the assessment and prediction of response to neoadjuvant chemotherapy (NAC) in terms of pathologic complete response (pCR) and post-NAC lymph node (LN) status in patients with node-positive invasive breast cancer, i.e., in patients with locally advanced breast cancer in whom the cancer has started to spread locally to the axilla.

The “early” prediction of breast cancer response to treatment using image-based phenotypes has gained interest in recent years with research focusing mainly on radiomics of MRI scans acquired up to after the first two cycles of NAC. We recently showed, using MR images of 141 women with 3 cm or greater breast cancers imaged at baseline in the publicly available I-SPY1 dataset, that the pretreatment most-enhancing tumor volume, an MR image-based radiomics feature, is predictive of recurrence-free survival.¹ We did not assess LNs in that study, however, and research by others has focused on the breast cancer itself as well. For example, looking only at pretreatment MRIs, another relatively large study found that within a multicenter independent validation cohort (186 patients), intratumoral spatial heterogeneity predicted recurrence-free survival in locally advanced breast cancer patients treated with NAC and that cancer aggressiveness was associated with larger poor perfusion subregions.^{2,3} Cain et al.⁴ found that radiomics of the tumor at pretreatment showed promise in the prediction of pCR in a subgroup of HER2-enriched and triple-negative cancers (151 patients in an independent test set of whom 28 achieved pCR). Another relatively large study of 117 patients also found promising performance of radiomics textural analysis of intratumoral and

*Address all correspondence to Karen Drukker, E-mail: kdrukker@uchicago.edu

peritumoral regions on pretreatment breast cancer dynamic contrast-enhanced (DCE) MRIs, with areas under the ROC curve up to 0.93 ± 0.018 for HER2-enriched and triple negative cancers combined in a separate threefold cross-validation analysis (47 patients).⁵ Likewise, in a multicenter study, a radiomics signature combined with independent clinicopathological risk factors achieved good performances in the prediction of pCR based on pretreatment multiparametric MRI in three external validation cohorts (with 99, 107, and 80 patients, respectively) with areas under the ROC curve up to 0.80.⁶

Several other radiomics studies using both pretreatment MRIs and those after the first cycle of NAC in the prediction of pCR have been reported: Dogan et al. found that DCE MRI kinetic parameters of tumors may have a role in predicting pCR in breast cancer.^{7,8} Sun et al. also looked at early prediction of pCR using MRIs acquired pretreatment and after the first cycle of NAC, and they obtained promising results in predicting pCR based on changes from pretreatment to after the first cycle of NAC in a validation sample of 34 patients.⁸ In a similar study, promising results were obtained in a small sample of 24 patients with locally advanced breast cancer.⁹ In another small sample of 35 patients diagnosed with stage II/III breast cancer 3 T DCE MR images acquired before and after the first cycle of NAC were analyzed, and it was found that analysis of tumor subregions yielded improved performance over whole tumor analysis.²

In other work focusing on predicting pCR using only MRIs obtained after the first cycle of NAC and comparing a pattern recognition-based method and a pharmacokinetic modeling approach in 35 patients, promising results were obtained for both methods with areas under the ROC curve ranging from 0.73 to 0.90 in their small patient sample.¹⁰ High performance in the prediction of treatment response was also reported by Tahmassebi et al.¹¹ in a small sample of 38 patients including all MRI scans acquired up to after the second cycle of NAC.

The goal of our current study is to investigate computer-extracted quantitative radiomics features of breast cancers and of metastatic axillary sentinel LNs for use in predicting pCR to NAC as well as in predicting the post-NAC LN status in patients with locally advanced breast cancer. In contrast to the work by others cited above, we use only MRIs acquired pre-NAC, consider the post-NAC LN status separately (similar to what we presented in a more preliminary analysis¹²), and include analysis of the axillary LNs. In a recent study, other researchers found that in head and neck cancer integrating tumor and nodal characteristics (and including pretreatment and mid-treatment exams) improved prediction of distant metastasis.¹³ It is important to note that, while achieving pCR in itself is only a moderate predictor of recurrence-free survival, patients with LN-positive invasive breast cancers could potentially benefit during treatment planning if it were possible to predict, before any treatment (using only pretreatment breast MRIs), which LNs and primary cancers would likely respond to NAC. For those patients, NAC could be more aggressive while potentially avoiding post-NAC axillary dissections and radiation, which are associated with significant morbidity as compared to treatment of the index lesion.

2 Methods

2.1 Dataset

Participants for our study were selected from a breast MRI database retrospectively collected at our institution. Inclusion

criteria were that the patient had confirmed invasive node-positive breast cancer, underwent MR imaging prior to any treatment, underwent NAC with known outcome in terms of pCR and post-NAC LN status (as determined at the time of surgery) and that the sentinel LN could be identified on the pretreatment MRI. Since the participants all had invasive node-positive breast cancer, achievement of pCR is generally defined as the absence of an invasive tumor component and the absence of nodal metastases at the time of surgery and axillary dissection; patients were defined as having achieved pCR if there was no remaining invasive cancer component and they were post-NAC LN-negative. LNs tend to respond to NAC before the primary tumor, however, and in our dataset this was the case for all participants. Hence, patients who did not achieve pCR could be divided into two (rather than three) groups in our cohort: (i) those in whom axillary metastases could no longer be identified post-NAC even though an invasive tumor component remained (no pCR but post-NAC LN-negative) and (ii) those in whom invasive cancer cells remained in both tumor and LNs at the time of surgery (no pCR and post-NAC LN-positive). There were no patients in whom the post-NAC LNs were positive but no invasive component of the index cancer remained. A single tumor (the index cancer) and a single LN (the sentinel axillary LN) were analyzed for each participant.

MR images were acquired between March 2009 and June 2017. Using Philips equipment (Philips Achieva, Koninklijke Philips, Eindhoven, the Netherlands) at a magnet field strength of either 1.5 T or 3.0 T, and we used only the DCE sequences (Table 1) between March 2009 and June 2017. Images were acquired prior to the injection of the contrast agent (gadolinium) and at time-intervals of ~ 1 min postcontrast agent injection with on average five postcontrast acquisitions (time-points). Only MRI exams acquired prior to the initiation of neoadjuvant chemotherapy were analyzed (Fig. 1).

Given that two magnet strengths were used in the acquisition of the images for our dataset, we used subsets by magnet strength in some of our analyses. The tumor sizes were similar for those imaged at 1.5 T and 3.0 T, but several extremely large LNs (>5 cm) were imaged at 1.5 T but none at 3.0 T (Table 1). Hence, these subsets were size-matched by pre-NAC LN size and contained the same number of cases acquired at 1.5 T as at 3.0 T (Table 1). The tumor size distributions remained similar by magnet strength after size-matching by LN size (Table 1). The hormone receptor status of the invasive breast cancer was available for all but four patients (Tables 1 and 2). For only about half of the patients detailed information on the administered NAC regimen was available for this project. At our institution, breast cancer NAC treatment is based on the American Society of Clinical Oncology and National Comprehensive Cancer Network guidelines and depends on the hormone receptor subtype. The most common NAC regimen followed for estrogen and/or progesterone-positive but HER2-negative cancers was cyclophosphamide, doxorubicin, pertuzamab, and taxol. For HER2-enriched cancers, the common regimens were (1) docetaxel with carboplatin, and concurrent HER2-targeted therapeutics trastuzumab with or without pertuzamab, (2) paclitaxel with carboplatin and concurrent HER2-targeted therapeutics trastuzumab with or without pertuzamab, and (3) anthracycline-based regimens for patients who lacked cardiac risk factors. The most common regimen followed for triple negative cancers was carboplatin, gemcitabine, and taxol.

Table 1 Description of the 158 women with invasive node-positive breast cancer imaged with DCE-MRI and subsets used in analyses (see “statistical analysis” section) where subsets A (acquired at 1.5 T) and B (acquired at 3 T) were size-matched by LN size and had comparable tumor size distributions.

Study participants		Complete dataset, <i>N</i> = 158 number of women (%)	Subset A, <i>N</i> = 61 number of women (%)	Subset B, <i>N</i> = 61 number of women (%)
Magnet strength	1.5 T	90 (57.0%)	61 (100%)	0 (0.0%)
	3.0 T	67 (42.4%)	0 (0.0%)	61 (100%)
	Unknown	1 (0.6%)	0 (0.0%)	0 (0.0%)
Age	≤40	28 (17.7%)	14 (23.0%)	7 (11.5%)
	>40 but ≤45	18 (11.4%)	4 (6.6%)	9 (14.8%)
	>45 but ≤50	28 (17.7%)	7 (11.5%)	16 (26.2%)
	>50 but ≤55	24 (15.2%)	12 (19.7%)	5 (8.2%)
	>55 but ≤65	29 (18.4%)	11 (18.0%)	12 (19.7%)
	>65 but ≤75	17 (10.8%)	5 (8.2%)	8 (13.1%)
	>75	8 (5.1%)	4 (6.6%)	3 (4.9%)
	Unknown	6 (3.8%)	4 (6.6%)	1 (1.6%)
	pCR	Yes	19 (12.0%)	6 (9.8%)
No		139 (88.0%)	55 (90.2%)	52 (85.2%)
Post-NAC LN status	Negative	42 (26.6%)	18 (29.5%)	18 (29.5%)
	Metastatic	116 (73.4%)	43 (70.5%)	43 (70.5%)
Size of pre-NAC invasive tumor (mm) ^a	≤5	0 (0.0%)	0 (0.0%)	0 (0.0%)
	>5 but ≤10	1 (0.6%)	0 (0.0%)	1 (1.6%)
	>10 but ≤20	11 (7.0%)	4 (6.6%)	5 (8.2%)
	>20 but ≤50	73 (46.2%)	29 (47.5%)	28 (45.9%)
	>50	73 (46.2%)	28 (45.9%)	27 (44.3%)
Size of pre-NAC metastatic LN (mm) ^a	≤5	0 (0.0%)	0 (0.0%)	0 (0.0%)
	>5 but ≤10	3 (1.9%)	1 (1.6%)	1 (1.6%)
	>10 but ≤20	49 (31.0%)	22 (36.1%)	21 (34.4%)
	>20 but ≤50	93 (58.9%)	38 (62.3%)	39 (63.9%)
	>50	13 (8.2%)	0 (0.0%)	0 (0.0%)
Hormone receptor status	HR-positive HER2-negative ^b	77 (48.7%)	32 (52.5%)	26 (42.6%)
	HER2-enriched	39 (24.7%)	15 (24.6%)	19 (31.1%)
	Triple-negative	38 (24.1%)	11 (18.0%)	15 (24.6%)
	Unknown	4 (2.5%)	3 (4.9%)	1 (1.6%)

^aMaximum linear size determined on imaging (DCE-MRI).

^bHR = hormone receptor (estrogen and/or progesterone), HER2 = human epidermal growth factor receptor 2.

2.2 Radiomics Method

Lesions (both tumors and LNs) were automatically segmented in the pre-NAC DCE-MRIs after manual localization of a seed-

point at the approximate lesion center by a breast imaging expert with over 22 years of experience and calculated from a bounding box drawn to enclose the entire lesion. Lesion segmentation was performed in four-dimensions (4-D) [three-dimensional (3-D)

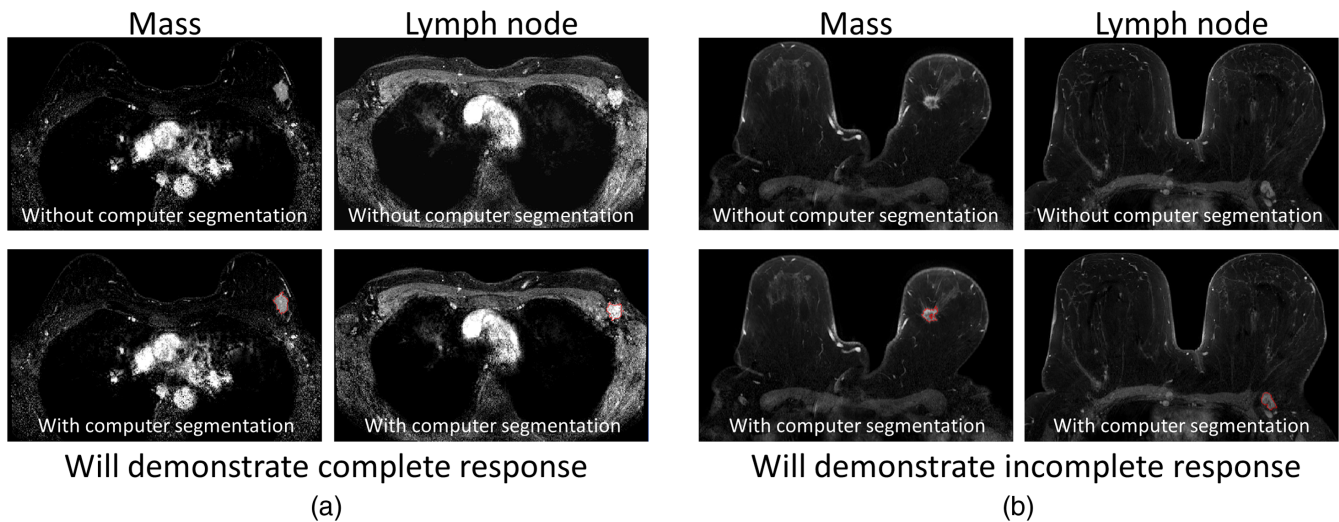


Fig. 1 Example pre-NAC MR image slices acquired at 3 T: (a) invasive breast cancer and LN positive for metastasis in a woman for whom NAC achieved pCR and (b) in a woman for whom NAC did not achieve pCR, respectively, (top) without and (bottom) with outlines of the computer segmentation outlines. Subtraction images (first postcontrast minus precontrast acquisitions) are shown. Note that the computer segmentations were derived in 4-D, i.e., using all the 3-D MR images for all acquisition times in a DCE MRI protocol.

Table 2 Treatment response by hormone receptor subgroup of the imaged node-positive invasive breast cancers for the entire dataset.

		Hormone receptor status			
		HR-positive HER2-negative ^a , N = 75	HER2-enriched, N = 39	Triple negative, N = 42	Unknown, N = 2
Complete dataset, N = 158					
pCR	Yes	3 (3.9%)	8 (20.5%)	7 (18.4%)	1 (25%)
	No	74 (96.1%)	31 (79.5%)	31 (81.6%)	3 (75%)
Post-NAC LN status	Negative	8 (10.4%)	18 (46.2%)	15 (39.5%)	1 (25%)
	Metastatic	69 (89.6%)	21 (53.8%)	23 (60.5%)	3 (75%)

^aHR = hormone receptor (estrogen and/or progesterone), HER2 = human epidermal growth factor receptor 2.

space plus DCE-acquisition time].¹⁴ Forty-nine radiomics features were extracted pertaining to seven categories describing (i) lesion size (three features), (ii) shape/geometry (three features), (iii) margin/morphology (three features), (iv) enhancement texture (14 features), (v) kinetics (10 features), (vi) variance kinetics (four features), and (vii) statistics or gray level histogram-based characteristics (12 features) (Table 3).^{15–19} Feature extraction was performed in 3-D and the enhancement texture features were calculated using the first postcontrast time-point (the MR image acquired about 1-min postcontrast-agent injection). Enhancement and kinetics-based features were extracted only from the most-enhancing tumor (or LN) regions, extracted using a second fuzzy c-means-based method within the previously segmented tumor or LN region only.¹⁶ The most-enhancing regions were used because of their proven merit in our previous studies in breast lesion classification^{15–19} as well as in prediction of breast cancer recurrence.¹ It is important to note that both segmentation and feature extraction were completely automated apart from the initial seed-point localization.

2.3 Statistical Analysis

2.3.1 Assessment of potential robustness of features with respect to magnet field strength

Imaged tumors and axillary LNs were analyzed separately. Since the images in our dataset were acquired at two magnet strengths, the first step was to determine which of the radiomics features (Table 3) were potentially robust with respect to magnet strength. For this purpose, we used the LN size-matched subsets A and B, acquired at 1.5 T and 3.0 T, respectively (Table 1), in order to allow for a “fair” comparison of feature value distributions by magnet field strength. The Mann–Whitney U-test was used with the null-hypothesis that samples were selected from populations having the same distribution. Features for which the distributions demonstrated a statistically significant difference ($p < 0.05$) in their distribution by magnet strength (subset A versus subset B, Table 1) were eliminated from further analysis. That is, we considered a feature “potentially robust” with respect to magnet field strength—and hence suitable for inclusion in

Table 3 Radiomics features by category extracted from the pre-NAC MRIs for the invasive cancers and for the metastatic axillary sentinel LNs. ¹⁵⁻¹⁹

Category	Feature name (unit)	Additional description	Label
Size	Volume (mm ³) [equivalent to effective diameter (mm)]	—	S ₁
	Surface area (mm ²)	—	S ₂
	Maximum diameter (mm)	Maximum distance between any two voxels in the lesion	S ₃
Shape/geometry	Sphericity	Resemblance of lesion shape to a sphere	G ₁
	Irregularity	Deviation of lesion surface from that of a sphere	G ₂
	Surface area-to-volume ratio (1/mm)	—	G ₃
Morphology	Margin sharpness	Mean of the image gradient at the lesion margin	M ₁
	Variance of margin sharpness	Variance of the image gradient at the lesion margin	M ₂
	Variance of radial gradient histogram	Degree to which the enhancement structure extends in a radial pattern originating from the center of the lesion	M ₃
Enhancement texture ^a	Angular second moment (energy)	Image homogeneity	T ₁
	Contrast	Local image variations	T ₂
	Correlation	Image linearity	T ₃
	Entropy	Randomness of the gray-levels	T ₄
	Sum of squares (variance)	Spread in the gray-level distribution	T ₅
	Difference entropy	Randomness of the difference of neighboring voxels' gray-levels	T ₆
	Difference variance	Variations of difference of gray-levels between voxel-pairs	T ₇
	Inverse difference moment	Image homogeneity	T ₈
	Sum average	Overall brightness	T ₉
	Sum entropy	Randomness of the sum of gray-levels of neighboring voxels	T ₁₀
	Sum variance	Spread in the sum of the gray-levels of voxel-pairs distribution	T ₁₁
	Information measure of correlation 1	Nonlinear gray-level dependence	T ₁₂
	Information measure of correlation 2	Nonlinear gray-level dependence	T ₁₃
Maximum correlation coefficient	Nonlinear gray-level dependence	T ₁₄	
Kinetic curve assessment	Maximum enhancement	Maximum contrast enhancement	K ₁
	Time to peak (s)	Time at which the maximum enhancement occurs	K ₂
	Uptake rate (1/s)	Uptake speed of the contrast enhancement	K ₃
	Washout rate (1/s)	Washout speed of the contrast enhancement	K ₄
	Curve shape index	Difference between late and early enhancement	K ₅
	Enhancement at first postcontrast time-point	Enhancement at first postcontrast time-point	K ₆
	Signal enhancement ratio	Ratio of initial enhancement to overall enhancement	K ₇
	Volume of most enhancing voxels (mm ³)	Volume of the most enhancing voxels	K ₈
	Total rate variation (1/s ²)	How rapidly the contrast will enter and exit from the lesion	K ₉
	Normalized total rate variation (1/s ²)	How rapidly the contrast will enter and exit from the lesion	K ₁₀

Table 3 (Continued).

Category	Feature name (unit)	Additional description	Label
Enhancement-variance kinetics	Maximum variance of enhancement	Maximum spatial variance of contrast enhancement over time	E ₁
	Time to peak at maximum variance (s)	Time at which the maximum variance occurs	E ₂
	Enhancement variance increasing rate (1/s)	Rate of increase of the enhancement-variance during uptake	E ₃
	Enhancement variance decreasing rate (1/s)	Rate of decrease of the enhancement-variance during washout	E ₄
Statistics	Mean voxel value within lesion precontrast injection	Average brightness (no contrast agent)	B ₁
	Mean voxel value within lesion first time-point postcontrast injection	Average brightness (~1 min after contrast-agent injection)	B ₂
	Standard deviation of voxel value distribution within lesion precontrast injection	Spread in brightness distribution (no contrast-agent)	B ₃
	Standard deviation of voxel value distribution within lesion first time-point postcontrast injection	Spread in brightness distribution (~1 min after contrast-agent injection)	B ₄
	Maximum voxel value within lesion precontrast injection	—	B ₅
	Maximum voxel value within lesion first time-point postcontrast injection	—	B ₆
	Minimum voxel value within lesion pre-contrast injection	—	B ₇
	Minimum voxel value within lesion first time-point post-contrast injection	—	B ₈
	Kurtosis of voxel value distribution within lesion precontrast injection	Sharpness of the peak of the brightness distribution (no contrast-agent)	B ₉
	Kurtosis of voxel value distribution within lesion first time-point postcontrast injection	Sharpness of the peak of the brightness distribution (~1 min after contrast-agent injection)	B ₁₀
	Skewness of voxel value distribution within lesion precontrast injection	Asymmetry of brightness distribution (no contrast-agent)	B ₁₁
	Skewness of voxel value distribution within lesion first time-point postcontrast injection	Asymmetry of brightness distribution (1 min after contrast-agent injection)	B ₁₂

^aEnhancement texture features were calculated using the first DCE-MRI acquisition after injection of the contrast agent (~1 min after contrast agent injection in our protocol).

subsequent analyses—when we failed to reject the null-hypothesis ($p > 0.05$) for the distributions by field strength for tumors and for the distributions by field strength for LNs. We did not correct p -values for multiple comparisons here in order to avoid labeling more and more features as “potentially robust” with respect to magnet strength.

2.3.2 Prediction of response by individual features

We investigated two end-points: (1) the pre-NAC prediction of pCR and (2) the pre-NAC prediction of the post-NAC LN status (negative versus metastatic). For these two end-points, “responders” were defined as those demonstrating pCR and

those demonstrating post-NAC negative LNs, respectively. Tumors and axillary LNs were again analyzed separately.

No classifier; bootstrap resampling. The ability of individual features on their own to distinguish between future “responders” and “non-responders,” i.e., predicting pCR or post-NAC LN status, was assessed using bootstrap resampling of the data (1000 samples). ROC analysis²⁰ was used to assess classification performance with the area under the ROC curve (AUC) as performance metric. Estimates for the 95% confidence intervals and p -values for superiority with respect to random guessing (AUC = 0.5) were obtained from the bootstrap samples. This analysis was performed for the entire dataset, subsets A + B combined, and by a hormone receptor subtype of the

invasive cancer: hormone receptor positive (estrogen and/or progesterone positive and HER2-negative), HER2-enriched, and triple negative (Tables 1 and 2). In the analysis by a hormone receptor subtype, ROC analysis was only used to assess the prediction of post-NAC LN status, not the prediction of pCR due to the very limited number of complete responders (e.g., three in the hormone-receptor positive cohort) in addition to the modest number of cases in the hormone receptor subgroups (Table 2).

Linear discriminant classifier; 632+ bootstrap training/testing. Classification performance was subsequently assessed for individual features in combination with a linear discriminant analysis classifier (with as input an individual feature) in a 632+ bootstrap training/testing paradigm (1000 iterations).²¹ ROC analysis²⁰ was again used to assess classification performance with the AUC as performance metric and estimates for the 95% confidence intervals and *p*-values for superiority with respect to random guessing obtained from the bootstrap samples (and corrected for bias according to the 632+ bootstrap approach). This analysis was performed for the entire dataset and for subsets A + B combined. Analysis by a hormone receptor subtype was not performed because sample sizes were too limited to perform reliable classifier training and testing (Table 2).

In summary, first features were assessed for potential robustness with respect to magnet field strength, then several subsets of the data were used to assess classification performance in the

task of predicting response to treatment (in terms of pCR and post-NAC LN status) (Table 4). We corrected *p*-values for multiple comparisons using Holm–Bonferroni²² assessing cancer and LN features separately. Features with a corrected *p*-value <0.05 were defined as outperforming random guessing. Note that the gold standard “truth” was determined at pathology, i.e., after surgery, rather than based on imaging. Also note that only single radiomics features, not feature combinations, were assessed—even in the analyses in which a classifier was used—and that all features were extracted from pre-NAC MRIs, i.e., before any treatment (Fig. 1).

3 Results

3.1 Assessment of Potential Robustness of Features with Respect to Magnet Field Strength

Many of the features demonstrated a statistically significant difference (*p* < 0.05) in their distribution by magnet strength. For 18 features, we were unable to reject the null-hypothesis that samples acquired at the two magnet strengths were drawn from the same distributions and thus were considered “potentially robust” with respect to magnet strength (Fig. 2). Only these features were included in further analyses and consequently the Holm–Bonferroni correction was applied to *p*-values for 18 comparisons throughout.

Table 4 Overview of the different analyses performed, data subset(s) used, and performance assessment methods (LDA = linear discriminant analysis).

Subset(s) used	“Potential robustness” assessment Comparison of feature value distributions by field strength (U-test)	Prediction of response			
		No classifier, bootstrap resampling (ROC)		LDA classifier, 632+ bootstrap training/testing (ROC)	
		pCR	post-NAC LN-status	pCR	post-NAC LN-status
Subset A (1.5 T) versus subset B (3.0 T) ^a	✓	N/A	N/A	N/A	N/A
Entire dataset ^a	N/A	✓	✓	✓	✓
Subsets A + B combined ^a	N/A	✓	✓	✓	✓
HR-positive only ^b	N/A	—	✓	—	—
HER2-enriched only ^b	N/A	—	✓	—	—
Triple-negative only ^b	N/A	—	✓	—	—

^aTable 1.

^bTable 2.

Features “potentially robust” with respect to the magnet field strength of acquisition Subset A (1.5T) vs. B (3.0T)



Fig. 2 Features for which we failed to reject the null-hypothesis for both tumors and LNs that the samples (obtained from images acquired at 1.5 T and 3.0 T, respectively) were drawn from populations with the same distribution, i.e., features “potentially robust” with respect to field strength within our dataset, shown in color and indicated with arrows.

3.1.1 Prediction of Response by Individual Features

No classifier; bootstrap resampling. In the pre-NAC prediction of pCR and the pre-NAC prediction of the post-NAC LN status, a single tumor feature and several LN features demonstrated initial promise but after correcting p -values for multiple comparisons, we failed to prove superiority to random guessing for any tumor feature. On the other hand, seven pre-NAC LN features were predictive of pCR when considering the entire dataset (Fig. 3) with corrected p -values in ascending order of 0.015, 0.016, 0.017, 0.018, 0.024, 0.026, and 0.028, respectively. A larger effective diameter of the LNs (feature S_1), a smaller surface area to volume ratio (G_3), and a more inhomogeneous appearance (the statistics features) were predictive of a positive outcome of NAC in terms of pCR. Even though the AUC values for the pre-NAC LN size-matched subsets A + B combined appeared similar to those for the entire dataset, fewer remained statistically significantly different from random guessing due to the smaller sample size (Fig. 3). In the prediction of post-NAC LN status, five statistics features performed significantly better than random guessing within the entire dataset (Fig. 3) with corrected p -values in ascending order of 0.015, 0.016, 0.017, 0.018, and 0.028, respectively. Again, fewer features remained predictive when considering the smaller sample of subsets A + B combined. It is interesting to note that, in the prediction of post-NAC LN status, only statistics features describing the pre-NAC imaged LNs appeared to be useful while the nodal size and geometry did not appear to be predictive in contrast to what was observed in the prediction of pCR.

In the pre-NAC prediction of post-NAC LN status by hormone receptor subgroup of the primary cancer (Table 2), only

three features outperformed random guessing after correcting the p -values for multiple comparisons (Fig. 3), and they all pertained to the pre-NAC metastatic LNs: two nodal statistics features for the hormone receptor-positive subgroup and a single nodal statistics feature for the HER2-enriched subgroup; for the hormone receptor-positive subgroup, the minimum precontrast (B_7) and kurtosis precontrast (B_9) within the nodes were predictive with AUC values of 0.77 [0.62 to 0.89] and 0.77 [0.63; 0.88] (corrected p -values of 0.034 and 0.036, respectively). For the HER2-enriched subgroup, the minimum postcontrast (B_8) within the nodes was predictive with an AUC value of 0.78 [0.62; 0.92] (corrected p -value = 0.018). For the triple negative subgroup, we failed to find any features predictive of response even though the total rate variation (K_9 , a kinetics feature) and the variance enhancement decreasing rate (E_4 , an enhancement variance feature) appeared somewhat promising with AUC values of 0.67 [0.47; 0.85] and 0.66 [0.48; 0.83], respectively. But since the 95% confidence intervals for AUC for these features included 0.5, we failed to prove superiority to random guessing.

Linear discriminant classifier; 632+ bootstrap training/testing. While overall the AUC values using the 632+ bootstrap classifier training/testing approach were very similar to those found in bootstrap resampling without a classifier, the 95% confidence intervals were slightly wider and perhaps a bit more pessimistic, which resulted in fewer features outperforming random guessing in the prediction of response, especially after correcting the p -values for multiple comparisons (Table 5). In the prediction of pCR, two LN features—surface area-to-volume ratio (G_3) and the maximum at the first post-contrast acquisition (B_6)—proved superior to random guessing

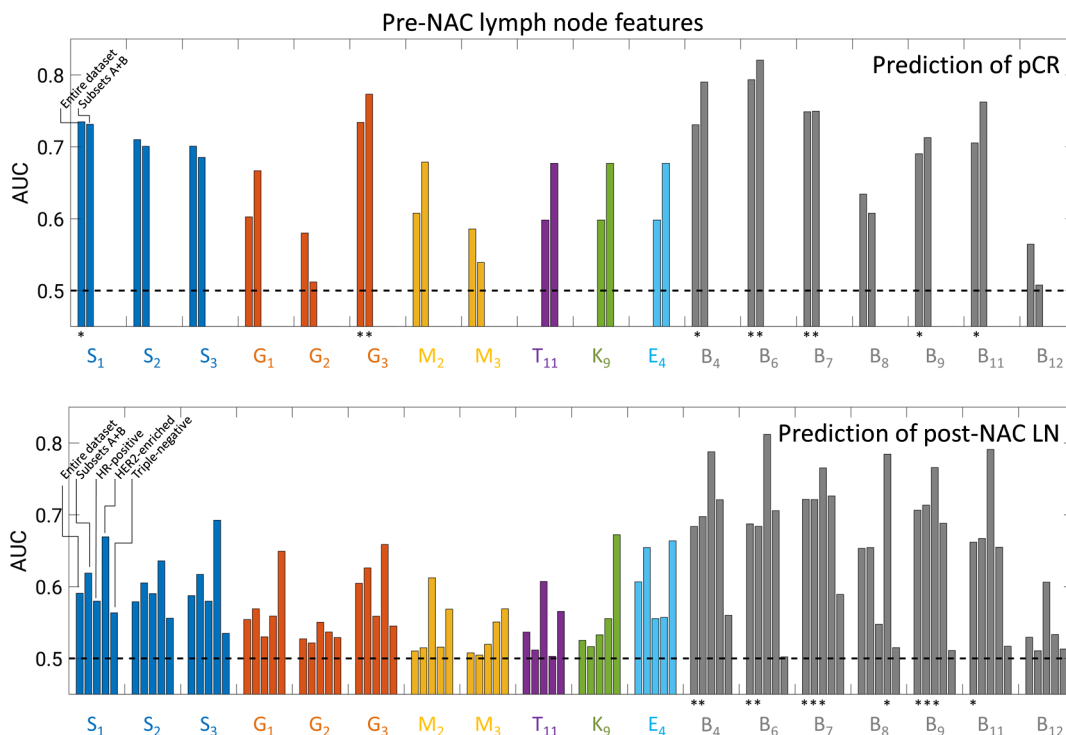


Fig. 3 Areas under the ROC curve (no classifier and bootstrap resampling) of individual pre-NAC LN features in the tasks of predicting pCR and post-NAC LN status, respectively, for the entire dataset, subsets A + B combined, HR-positive only, HER2-enriched only, and triple-negative subgroups as indicated. An asterisk indicates that performance remained significantly better than random guessing after correcting for multiple comparisons (18 comparison, see the text for p -values). Error bars are not shown for clarity.

Table 5 Areas under the ROC curve (linear discriminant analysis classifier, 632+ bootstrap training/testing) in the tasks of predicting pCR and post-NAC LN status, respectively, for the entire dataset and subsets A + B combined: AUC values for features that outperformed random guessing (AUC = 0.5) before correcting statistical significance for multiple comparisons (regular font) and AUC values for features that remained significantly better than random guessing after correcting for multiple comparisons (18 comparisons, see text for p -values) (bold font).

Feature	Pre-NAC LN feature in prediction of pCR		Pre-NAC LN feature in prediction of post-NAC LN-status	
	Entire dataset ($N = 158$)	Subsets A + B ($N = 122$)	Entire dataset ($N = 158$)	Subsets A + B ($N = 122$)
S ₁	0.73 [0.58; 0.82]	—	—	—
S ₂	0.71 [0.51; 0.79]	—	—	—
S ₃	—	—	—	—
G ₁	—	—	—	—
G ₂	—	—	—	—
G ₃	0.73 [0.62; 0.80]	0.77 [0.61; 0.85]	—	—
M ₂	—	—	—	—
M ₃	—	—	—	—
T ₁₁	—	—	—	—
K ₉	—	—	—	—
E ₄	—	—	—	—
B ₄	0.73 [0.53; 0.81]	0.78 [0.60; 0.87]	0.68 [0.51; 0.76]	0.69 [0.55; 0.76]
B ₆	0.79 [0.69; 0.85]	0.82 [0.70; 0.88]	0.68 [0.51; 0.77]	—
B ₇	0.80 [0.61; 0.82]	0.74 [0.55; 0.83]	0.71 [0.62; 0.79]	0.72 [0.62; 0.77]
B ₈	—	—	—	—
B ₉	0.69 [0.53; 0.77]	0.71 [0.51; 0.80]	0.70 [0.59; 0.78]	0.71 [0.60; 0.77]
B ₁₁	—	0.75 [0.52; 0.84]	0.66 [0.53; 0.75]	0.66 [0.52; 0.74]
B ₁₂	—	—	—	—

(corrected p -values 0.044 and 0.048, respectively) in the analysis of the entire dataset. For subsets A + B combined, only B_6 was proven to be superior to random guessing (Table 5). In the prediction of post-NAC LN status, two different LN features emerged as being able to predict response (both for the entire dataset and for subsets A + B combined): minimum precontrast (B_7) and kurtosis precontrast (B_9) with adjusted p -values of 0.017 and 0.018 (Table 5).

4 Discussion

In the pretreatment prediction of response to NAC in patients with node-positive invasive breast cancer, features of the primary tumors imaged pre-NAC with DCE-MRI appeared to have limited usefulness even though we previously found some of these tumor features to be useful in predicting recurrence-free survival in a different patient cohort,¹ and several other researchers found radiomics tumor features to be useful in the prediction of pCR as detailed in the Introduction³⁻⁶ as well as in the prediction of the pre-NAC LN status in breast cancer patients.^{23,24} Features of pre-NAC imaged metastatic axillary sentinel LNs, on the other hand, demonstrated promise in the prediction of treatment response, with the areas under the ROC curve in the

bootstrap 632+ analyses up to 0.82 [0.70; 0.88] in the prediction of pCR and up to 0.72 [0.62; 0.77] in the prediction of post-NAC LN status. More compact, more inhomogeneous appearing pre-NAC metastatic axillary LNs were predictive of more successful NAC treatment in terms of pCR, and statistics features of the imaged metastatic LNs were most successful in predicting post-NAC LN status. In the analysis by hormone receptor subgroup, the prediction of response to treatment showed promise in the hormone receptor-positive and HER2-enriched subgroups but image features failed to be predictive for triple negative breast cancers even though the response rate for triple negative cancers and HER2-enriched cancers were similar in our dataset.

The fact that pretreatment tumor features extracted from MRI may not be informative in predicting treatment response was also found by Nilsen et al.²⁵ in a diffusion-weighted MRI study of 25 patients with locally advanced breast cancer. These authors found that the pretreatment tumor apparent diffusion coefficient failed to predict treatment response and that the increase in the apparent diffusion coefficient observed mid-way in the course of NAC failed to show correlation with tumor volume changes. Another larger study using diffusion-weighted

MRI in 164 breast cancer patients also found that the pretreatment apparent diffusion coefficient failed to be predictive of treatment response.²⁶ On the other hand, there was a statistically significant difference in the apparent diffusion coefficient for responders and nonresponders after the second cycle of NAC and also the change in the apparent diffusion coefficient over time was predictive of treatment response.²⁶ Interestingly, a study in 83 patients with locally advanced breast cancer found that the texture on PET was an independent predictor of pCR.²⁷ To our knowledge, our current study is the first investigating pre-NAC LN features in patients with locally advanced breast cancer for the prediction of response to NAC, both in terms of pCR and in terms of post-NAC LN status. Our analysis by a cancer hormone receptor subtype is of interest since it was shown that pCR predicts recurrence-free survival more effectively by cancer subtype.²⁸

Limitations of the current pilot study included the modest size of the dataset (however, the dataset was comparable in size to the I-SPY1 dataset), the imbalance of the dataset (few patients achieved pCR or post-NAC negative LNs), the unavailability of specific details of the NAC regimens for many patients, the unavailability of survival data for most patients, and the different magnet strengths used in MRI acquisition (1.5 T and 3.0 T). We ameliorated the latter by assessing only features that appeared unaffected by magnet strength. We are also currently expanding our investigation of the dependence of MRI radiomics mass features on magnet field strength.²⁹ One should note that the most-enhancing tumor volume, which was shown to be predictive of recurrence-free survival in prior work¹ using the publicly available I-SPY1 dataset^{30–32} (in which all MRIs were acquired at 1.5 T), was found to depend on magnet strength and was hence not included in the current analysis combining images acquired at 1.5 T and at 3.0 T. One should also note that pCR and post-NAC LN status are intermediate outcomes, not necessarily predictive of long-term recurrence-free, or overall, survival. For example, patients with hormone receptor-positive breast cancer (estrogen and/or progesterone positive and HER2-negative) generally have a good prognosis in spite of relatively weak response to NAC due to the success of surgery and adjuvant treatment with hormone therapy drugs.^{33,34} However, the ability to predict pCR and post-NAC LN status could positively impact treatment plans by identifying patients in whom NAC is likely to be successful and hence could be used more aggressively while subjecting fewer of these patients to unnecessary, and potentially harmful, procedures such as axillary dissection. In patients for whom NAC is identified as less likely to be successful, on the other hand, treatments other than NAC could be considered.

Future work will include the collection of a larger and longitudinal dataset including patient outcome and assessment of tumor and LN signatures (multiple features in combination with a classifier) as well as recurrence-free survival, building upon our current study and prior publication.¹

Disclosures

Karen Drukker receives royalties from Hologic. Maryellen Giger is a stockholder in Hologic Inc., is cofounder and equity holder in Quantitative Insights Inc. (now Qlarity Imaging), and receives royalties from Hologic Inc., General Electric Company, MEDIAN Technologies, Riverain Technologies LLC, Mitsubishi Corporation and Toshiba Corporation.

Acknowledgments

This project was funded in part by NIH Grant No. U01CA195564 in the QIN (Quantitative Imaging Network).

References

1. K. Drukker et al., "Most-enhancing tumor volume by MRI radiomics predicts recurrence-free survival 'early on' in neoadjuvant treatment of breast cancer," *Cancer Imaging* **18**(1), 12 (2018).
2. J. Wu et al., "Intratumor partitioning and texture analysis of dynamic contrast-enhanced (DCE)-MRI identifies relevant tumor subregions to predict pathological response of breast cancer to neoadjuvant chemotherapy," *J. Magn. Reson. Imaging* **44**(5), 1107–1115 (2016).
3. J. Wu et al., "Intratumor spatial heterogeneity at perfusion MR imaging predicts recurrence-free survival in locally advanced breast cancer treated with neoadjuvant chemotherapy," *Radiology* **288**(1), 26–35 (2018).
4. E. H. Cain et al., "Multivariate machine learning models for prediction of pathological response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set," *Breast Cancer Res. Treat.* **173**(2), 455–463 (2019).
5. N. M. Braman et al., "Intratumor and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI," *Breast Cancer Res.* **19**(1), 57 (2017).
6. Z. Liu et al., "Radiomics of multi-parametric MRI for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study," *Clin. Cancer Res.* **25**(12), 3538–3547 (2019).
7. B. E. Dogan et al., "Comparing the performances of magnetic resonance imaging size vs pharmacokinetic parameters to predict response to neoadjuvant chemotherapy and survival in patients with breast cancer," *Curr. Probl. Diagn. Radiol.* **48**(3), 235–240 (2018).
8. Y. S. Sun et al., "Predictive value of DCE-MRI for early evaluation of pathological complete response to neoadjuvant chemotherapy in resectable primary breast cancer: a single-center prospective study," *Breast* **30**, 80–86 (2016).
9. R. Johansen et al., "Predicting survival and early clinical response to primary chemotherapy for patients with locally advanced breast cancer using DCE-MRI," *J. Magn. Reson. Imaging* **29**(6), 1300–1307 (2009).
10. E. Kontopodis et al., "Investigating the role of model-based and model-free imaging biomarkers as early predictors of neoadjuvant breast cancer therapy outcome," *IEEE J. Biomed. Health Inform.* **23**(5), 1834–1843 (2019).
11. A. Tahmassebi et al., "Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients," *Invest. Radiol.* **54**(2), 110–117 (2019).
12. K. Drukker et al., "Breast MRI radiomics for the pre-treatment prediction of response to neoadjuvant chemotherapy in node-positive breast cancer patients," *Proc. SPIE* **10950**, 109502N (2019).
13. J. Wu et al., "Integrating tumor and nodal imaging characteristics at baseline and mid-treatment computed tomography scans to predict distant metastasis in oropharyngeal cancer treated with concurrent chemoradiotherapy," *Int. J. Radiat. Oncol. Biol. Phys.* **104**(4), 942–952 (2019).
14. W. Chen, M. L. Giger, and U. Bick, "A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images," *Acad. Radiol.* **13**(1), 63–72 (2006).
15. K. G. Gilhuijs, M. L. Giger, and U. Bick, "Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging," *Med. Phys.* **25**(9), 1647–1654 (1998).
16. W. Chen et al., "Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI," *Med. Phys.* **33**(8), 2878–2887 (2006).
17. W. Chen et al., "Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images," *Magn. Reson. Med.* **58**(3), 562–571 (2007).
18. W. Chen et al., "Computerized interpretation of breast MRI: investigation of enhancement-variance dynamics," *Med. Phys.* **31**(5), 1076–1082 (2004).

19. W. Chen et al., "Computerized assessment of breast lesion malignancy using DCE-MRI robustness study on two independent clinical datasets from two manufacturers," *Acad. Radiol.* **17**(7), 822–829 (2010).
20. C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.* **8**(4), 283–298 (1978).
21. B. Sahiner, H. P. Chan, and L. Hadjiiski, "Classifier performance prediction for computer-aided diagnosis using a limited dataset," *Med. Phys.* **35**(4), 1559–1570 (2008).
22. S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.* **6**(2), 65–70 (1979).
23. C. Liu et al., "Preoperative prediction of sentinel lymph node metastasis in breast cancer by radiomic signatures from dynamic contrast-enhanced MRI," *J. Magn. Reson. Imaging* **49**(1), 131–140 (2019).
24. J. Yang et al., "Preoperative prediction of axillary lymph node metastasis in breast cancer using mammography-based radiomics method," *Sci. Rep.* **9**(1), 4429 (2019).
25. L. Nilsen et al., "Diffusion-weighted magnetic resonance imaging for pretreatment prediction and monitoring of treatment response of patients with locally advanced breast cancer undergoing neoadjuvant chemotherapy," *Acta Oncol.* **49**(3), 354–360 (2010).
26. X. Y. Hu et al., "Diffusion-weighted MR imaging in prediction of response to neoadjuvant chemotherapy in patients with breast cancer," *Oncotarget* **8**(45), 79642–79649 (2017).
27. H. J. Yoon et al., "Predicting neo-adjuvant chemotherapy response and progression-free survival of locally advanced breast cancer using textural features of intratumoral heterogeneity on F-18 FDG PET/CT and diffusion-weighted MR imaging," *Breast J.* **25**(3), 373–380 (2018).
28. L. J. Esserman et al., "Pathologic complete response predicts recurrence-free survival more effectively by cancer subset: results from the I-SPY 1 TRIAL–CALGB 150007/150012, ACRIN 6657," *J. Clin. Oncol.* **30**(26), 3242–3249 (2012).
29. H. Whitney et al., "Robustness of radiomic breast features of benign lesions and luminal A cancers across MR magnet strengths," *Proc. SPIE* **10575**, 105750A (2018).
30. N. M. Hylton et al., "Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy—results from ACRIN 6657/I-SPY TRIAL," *Radiology* **263**(3), 663–672 (2012).
31. N. M. Hylton et al., "Neoadjuvant chemotherapy for breast cancer: functional tumor volume by MR imaging predicts recurrence-free survival—results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL," *Radiology* **279**(1), 44–55 (2016).
32. D. Newitt and N. Hylton, "[on behalf of the I-SPY 1 Network and ACRIN 6657 Trial Team. (2016). Multi-center breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials]," The Cancer Imaging Archive, <http://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK> (2016).
33. E. Munzone and M. Colleoni, "Optimal management of luminal breast cancer: how much endocrine therapy is long enough?" *Ther. Adv. Med. Oncol.* **10**, 175883591877743 (2018).
34. R. K. C. Ngan, "Management of hormone-receptor positive human epidermal receptor 2 negative advanced or metastatic breast cancers," *Ann. Transl. Med.* **6**(14), 284 (2018).

Karen Drukker PhD is a research associate professor of radiology at the University of Chicago and has been involved in radiomics-related research for almost two decades, mostly within breast imaging.

John Papaioannou is a computer scientist in the Department of Radiology at the University of Chicago. He holds a Master of Science in computer science (computer vision) from Northwestern University. He has expertise in medical imaging database structuring and implementation and has been instrumental in the computer-aided diagnosis and machine learning research in the Giger Lab at the university.

Kirti Kulkarni is the director of the Breast Imaging Fellowship program at the University of Chicago Medicine. She is an associate professor and physician-scientist in the Department of Radiology. Her clinical interests include creating awareness and early breast cancer diagnosis and treatment. Her research focuses on newer imaging tools to assess breast specimen margins intra-operatively.

Maryellen L. Giger is the A. N. professor of Radiology/Medical Physics at the University of Chicago and has been working for multiple decades on computer-aided diagnosis/computer vision/machine learning/deep learning in cancer diagnosis and management. Her research interests include understanding the role of quantitative radiomics and machine learning in personalized medicine.