

Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data

Wentian Guo
Hui Li
Yitan Zhu
Li Lan
Shengjie Yang
Karen Drukker
Elizabeth Morris
Elizabeth Burnside
Gary Whitman
Maryellen L. Giger
Yuan Ji
TCGA Breast Phenotype Research Group

Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data

Wentian Guo,^{a,d,†} Hui Li,^{b,†} Yitan Zhu,^{c,†} Li Lan,^b Shengjie Yang,^c Karen Drukker,^b Elizabeth Morris,^e Elizabeth Burnside,^f Gary Whitman,^g Maryellen L. Giger,^{b,*} Yuan Ji,^{a,c,*} and TCGA Breast Phenotype Research Group^h

^aUniversity of Chicago, Department of Public Health Sciences, 5841 South Maryland Avenue MC2000, Chicago, Illinois 60637, United States

^bUniversity of Chicago, Department of Radiology, 5841 South Maryland Avenue, Chicago, Illinois 60637, United States

^cNorthShore University Health System, Program of Computational Genomics & Medicine, 1001 University Place, Evanston, Illinois 60201, United States

^dFudan University, School of Public Health, 130 Dongan Road, Shanghai 200032, China

^eMemorial Sloan Kettering Cancer Center, Department of Radiology, 1275 York Avenue, New York, New York 10065, United States

^fUniversity of Wisconsin, School of Medicine and Public Health, Department of Radiology, E3/366 Clinical Science Center, 600 Highland Avenue, Madison, Wisconsin 53792-3252, United States

^gMD Anderson, 1515 Holcombe Boulevard, Houston, Texas 77030, United States

^h<https://wiki.cancerimagingarchive.net/display/Public/TCGA+Breast+Phenotype+Research+Group>

Abstract. Genomic and radiomic imaging profiles of invasive breast carcinomas from The Cancer Genome Atlas and The Cancer Imaging Archive were integrated and a comprehensive analysis was conducted to predict clinical outcomes using the radiogenomic features. Variable selection via LASSO and logistic regression were used to select the most-predictive radiogenomic features for the clinical phenotypes, including pathological stage, lymph node metastasis, and status of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Cross-validation with receiver operating characteristic (ROC) analysis was performed and the area under the ROC curve (AUC) was employed as the prediction metric. Higher AUCs were obtained in the prediction of pathological stage, ER, and PR status than for lymph node metastasis and HER2 status. Overall, the prediction performances by genomics alone, radiomics alone, and combined radiogenomics features showed statistically significant correlations with clinical outcomes; however, improvement on the prediction performance by combining genomics and radiomics data was not found to be statistically significant, most likely due to the small sample size of 91 cancer cases with 38 radiomic features and 144 genomic features. © 2015 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.2.4.041007]

Keywords: radiogenomics; invasive breast carcinoma; prediction of clinical outcomes; The Cancer Genome Atlas; The Cancer Imaging Archive.

Paper 15081SSRR received Apr. 13, 2015; accepted for publication Jul. 22, 2015; published online Sep. 23, 2015.

1 Introduction

Radiogenomics integrates genomic and radiomic imaging profiles and has become an increasingly important research direction, especially in cancer, due to its potential to improve disease diagnosis, prognosis, and treatment choice.^{1,2} Radiological imaging uses noninvasive procedures, such as x-ray computed tomography and magnetic resonance imaging (MRI), to assess the phenotype characteristics of tumor and such imaging is routinely used in clinical practice.^{3,4} Genomic profiling of tumor is usually obtained through invasive procedures, such as biopsy and surgery, providing direct observation on the molecular underpinnings of the tumor. The integration of these two different data modalities provides opportunities to investigate whether combining radiomics and genomics can achieve better prediction of tumor clinical types than using either alone.

Many clinical outcomes in oncology are closely related to cancer diagnosis, prognosis, and treatment planning. For invasive breast carcinoma, pathological stage and molecular receptor

status are important variables considered in clinical practice. Pathological stage is based on the T-N-M classification of tumors.⁵ T stage describes the size of the primary tumor and its invasion into the surrounding tissue; N stage evaluates the involvement of nearby lymph nodes; and M stage indicates distant metastasis of cancer. An overall pathological stage based on T, N, and M classifications is summarized for each cancer case. For the molecular receptor status of a patient, estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) are usually considered,⁶ and treatments specific to the status of these receptors have been developed. For example, trastuzumab and lapatinib are quite effective for treating HER2+ breast cancer patients.⁷ In summary, these clinical variables can stratify breast cancer patients into subgroups, with different disease severities, mechanisms, and treatment schemes.

Most of the existing research that integrates genomic data with radiomic imaging data are conducted to elucidate the correlations between genomic features and imaging features, as the latter is noninvasive and more inclusive of the entire tumor than the former.^{1,8–10} In this study, we explore the relationship

*Address all correspondence to: Maryellen L. Giger, E-mail: m-giger@uchicago.edu; Yuan Ji, E-mail: koeraser@gmail.com

[†]These authors contributed equally.

between integrative radiogenomic features and several important clinical variables including pathology stage, lymph node metastasis, and molecular receptor status. Innovatively, we assess whether combining genomic and radiomic features can improve the prediction of clinical outcomes.

Radiomic features and genomic features are two distinct types of measurements of tumor. Radiomic features are closely related to tumor phenotypes, while genomic features characterize the underlying genetic and molecular profile of a tumor. Each of the two feature types can be used for possible determination or prediction of tumor characteristics and progression. By combining the two different feature types, more accurate and complete understanding of a tumor might be achieved than when using each of them alone. Such improved understanding may further help improve disease diagnosis and prognosis, thus facilitating better clinical decisions on patient care.

We analyzed a unique radiogenomic dataset consisting of 91 breast cancer patients. In particular, we extracted genomic data and radiomic images of 91 patients from The Cancer Genome Atlas (TCGA)¹¹ and The Cancer Imaging Archive (TCIA)¹² projects of the U.S. National Cancer Institute, respectively. Two kinds of statistical analyses were performed. First, *t* tests were used to learn the relationship between clinical outcomes and individual radiogenomic features. Second, we used logistic regression with LASSO regularization¹³ to select genomic features and radiomic features predictive of clinical outcomes and to assess their prediction power.

2 Materials and Methods

2.1 Clinical Data

There were a total of 91 invasive breast carcinomas with both radiomic imaging profiles from TCIA and genomic data from TCGA. Clinical information of the 91 cases was downloaded from TCGA using TCGA-Assembler.¹⁴

All samples were primary tumors from female patients. The patients' average age was 53.6 years with a standard deviation of 11.5 years and a range of 29 to 82 years with a median of 53 years. Out of the 91 invasive breast cancer cases, 87% (79/91) were ductal carcinoma, 11% (10/91) were lobular carcinoma, and 2% (2/91) were mixed. Only one patient was listed as having death as an outcome.

Patient ER status is a binary response, with 77 patients being ER+ and 14 patients being ER-. Regarding PR status, 72 patients are PR+ and 19 patients are PR-. HER2 status was missing, equivocal, positive, and negative in 6, 22, 14, and 49 patients, respectively. Only the HER2+ and HER2- samples were used in the analysis. We studied the prediction of ER, PR, and HER2 status of patients using radiomic features alone, genomic features alone, and the combination of both types of features. All genomic features of the ERBB2 gene were excluded from the analysis of predicting HER2 status because ERBB2 and HER2 are two aliases of the same gene. Therefore, we excluded ERBB2 features to avoid adding obvious confounders to the predictive models.

We also studied the discrimination between 22 stage I tumors and 11 stage III tumors, since stage I and stage III represent less-aggressive and more-aggressive tumors, respectively. For the prediction of lymph node metastasis, one patient sample was excluded from the analysis, because the number of lymph nodes with metastasis was missing in it. We dichotomized the samples into two classes, which are with and without lymph

node metastasis, and the numbers of patients belonging to each class were 42 and 48, respectively.

2.2 Imaging Data and Radiomic Features

Dynamic contrast enhanced (DCE) MRI data of the 91 tumors were downloaded from TCIA. There were 108 MRI examinations available at the time of this study. In order to reduce the image acquisition variation, only the breast MRIs acquired on a GE scanner with 1.5 T magnet strength were analyzed (i.e., 93 cases). In addition, one case with missing images in the dynamic sequence and one case without genomic data were excluded from the study. The resulting 91 cases in the final dataset had been contributed by four institutions with examination dates ranging from 1999 to 2004.

MR images used in this study had been acquired with a standard double breast coil on a 1.5 T GE whole-body MRI system. Only T1-weighted DCE MRIs were used for the study. The imaging protocols included one pre- and three to five postcontrast images obtained using a T1-weighted three-dimensional spoiled gradient echo sequence with a gadolinium-based contrast agent.

Each MRI exam was independently reviewed by three experienced TCIA breast radiologists blinded to the outcome data. The primary tumor location on MRI was determined by consensus using the radiologists' annotated information on the images. This tumor location information was the only input for the quantitative image analysis of the breast tumor on MRI. Prior to the computer extraction of the various image phenotypes, the tumor was segmented on the MRI using the radiologist-indicated tumor center and a computational fuzzy c-means algorithm.¹⁵

Quantitative radiomics analysis was then conducted,^{16-26,27} yielding 38 radiomic features characterizing the size, shape, morphology, enhancement texture, kinetics, and variance kinetics of each tumor. These radiomic features can be sorted into six MRI phenotype categories: (1) size, giving the tumor dimensions, such as volume and surface area, (2) shape, characterizing the tumor geometry, such as sphericity and irregularity, (3) morphology, combining tumor shape and margin characteristics, such as spiculation and margin sharpness, (4) enhancement texture, characterizing tumor textural properties based on the gray-level co-occurrence matrix, such as energy, entropy, and contrast, (5) kinetic curve assessment, characterizing the physiological process of the uptake and washout nature of the contrast agent in a breast tumor during the dynamic imaging series, such as uptake rate, washout rate, and signal enhancement ratio, and (6) enhancement-variance kinetic features, characterizing the time course of the spatial variance of the enhancement within a breast tumor, such as variance increase rate and variance decrease rate. Information about the radiomic features, including feature name, label, description, and category, is listed in Table 1.

Figure 1 shows the images of two tumor cases with the information of some clinical variables and the values of some radiomic features.

2.3 Genomic Data

Genes were selected based on two recently published papers^{28,29} which explored genes involved in breast cancer. These two papers discussed genes that are expected to influence the germination and progress of breast cancer. Genomic data of these genes in the 91 tumors were downloaded from TCGA using TCGA-Assembler¹⁴ to obtain three types of genomic features:

Table 1 Information about the 38 radiomic features.

Feature category	Label	Name	Description
Size features	S1	Lesion volume (mm ³)	Volume of lesion
	S2	Effective diameter (mm)	Diameter of a sphere with the same volume as the lesion
	S3	Surface area (mm ²)	Lesion surface area
	S4	Maximum linear size (mm)	Maximum distance between any two voxels in the lesion
Shape features	G1	Sphericity	Similarity of the lesion shape to a sphere
	G2	Irregularity	Deviation of the lesion surface from the surface of a sphere
	G3	Surface-to-volume ratio (1/mm)	Ratio of surface area to volume
Morphological features	M1	Margin sharpness	Mean of the image gradient at the lesion margin
	M2	Variance of margin sharpness	Variance of the image gradient at the lesion margin
	M3	Variance of radial gradient histogram	Indicates how well the enhancement structure in a lesion extends in a radial pattern originating from the center of the lesion
Enhancement textures	T1	Contrast	Measure of local image variations
	T2	Correlation	Measure of image linearity
	T3	Difference entropy	Measure of the randomness of the difference of neighboring voxels' gray levels
	T4	Difference variance	Measure of variations of difference of gray levels between voxel pairs
	T5	Angular second moment (energy)	Measure of image homogeneity
	T6	Entropy	Measure of the randomness of the gray levels
	T7	Inverse difference moment	Measure of the image homogeneity
	T8	Information measure of correlation 1	Measure of nonlinear gray-level dependence
	T9	Information measure of correlation 2	Measure of nonlinear gray-level dependence
	T10	Maximum correlation coefficient	Measure of nonlinear gray-level dependence
	T11	Sum average	Measure of the overall image brightness
	T12	Sum entropy	Measure of the randomness of the sum of gray levels of neighboring voxels
	T13	Sum variance	Measure of the spread in the sum of the gray levels of voxel-pairs distribution
	T14	Sum of squares (variance)	Measure of the spread in the gray-level distribution

Table 1 (Continued).

Feature category	Label	Name	Description
Kinetic curve assessments	K1	Maximum enhancement	Maximum contrast enhancement
	K2	Time to peak (s)	Time at which the maximum enhancement occurs
	K3	Uptake rate (1/s)	Uptake speed of the contrast enhancement
	K4	Washout rate (1/s)	Washout speed of the contrast enhancement
	K5	Curve shape index	Difference between late and early enhancement
	K6	Enhancement at first postcontrast time point	Enhancement at first postcontrast time point
	K7	Signal enhancement ratio	Ratio of initial enhancement to overall enhancement
	K8	Volume of most enhancing voxels (mm ³)	Volume of the most enhancing voxels
	K9	Total rate variation (1/s ²)	Measures how rapidly the contrast will enter and exit from the lesion
	K10	Normalized total rate variation (1/s ²)	Measures how rapidly the contrast will enter and exit from the lesion
Enhancement-variance kinetics	E1	Maximum variance of enhancement	Maximum spatial variance of contrast enhancement over time
	E2	Time to peak at maximum variance (s)	Time at which the maximum variance occurs
	E3	Enhancement variance increasing rate (1/s)	Rate of increase of the enhancement variance during uptake
	E4	Enhancement variance decreasing rate (1/s)	Rate of decrease of the enhancement-variance during washout

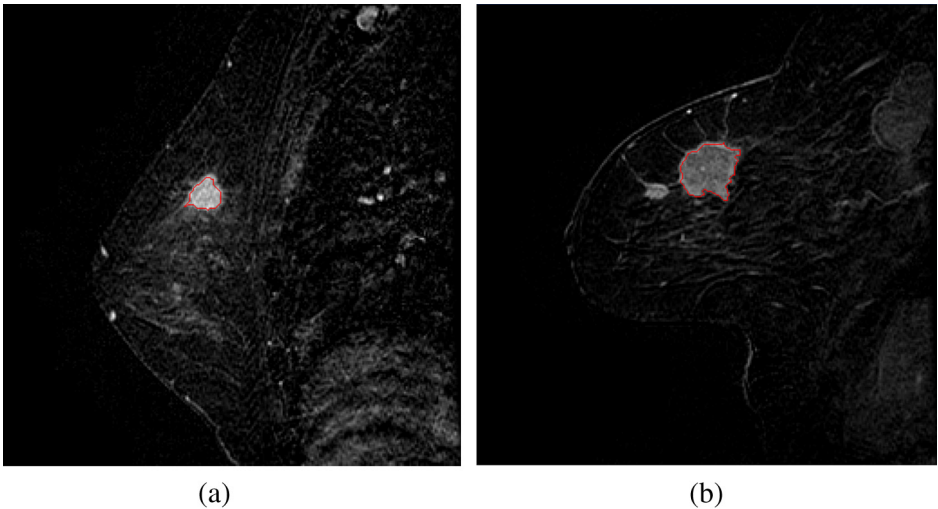


Fig. 1 Example cases including segmentation outlines obtained by the computational segmentation algorithm. (a) A Luminal A tumor from a 52-year-old female that is estrogen receptor (ER)-positive, progesterone receptor (PR)-positive, human epidermal growth factor receptor 2 (HER2)-negative, stage II, and without lymph node metastasis. The effective diameter, shape irregularity, and angular second moment (energy) of this tumor are 13.6 mm, 0.49, and 0.00185, respectively. (b) A HER2-enriched tumor from a 79-year-old female that is ER-negative, PR-negative, HER2-positive, stage III, and with lymph node metastasis. The effective diameter, shape irregularity, and angular second moment (energy) of this tumor are 26.4 mm, 0.47, and 0.00192, respectively.

copy number (CN), gene expression (GE), and DNA methylation (ME). Protein expressions were not considered as they were missing in a large portion of the samples. More explanations of CN, GE, and ME features are included in the [Appendix](#). For gene expression data, we used the normalized read counts of RNA-seq data, which were generated by TCGA using the Illumina HiSeq 2000 system and processed using the MapSplice genome alignment algorithm³⁰ and the RSEM gene expression estimation algorithm.³¹ TCGA used the Affymetrix Genome-Wide Human SNP Array 6.0 and the circular binary segmentation algorithm³² to obtain gene CNs. ME was measured using Infinium HumanMethylation450 BeadChip. For CN, GE, and ME, TCGA did not have all three features for every sample. Twenty-six features did not have measurements in 29 patients, which were nearly one third of the patients. Thus, these features were removed from analysis. One patient did not have data for four methylation features, including PTEN(ME), TP53(ME), AFF2(ME), and ATM(ME). These missing values were imputed using the sample mean of the feature across other patients whose data were present.

In the end, a genomic dataset was obtained with 144 genomic features for 70 genes, including 70 gene expression features, 70 CN features, and 4 methylation features. The full list of genes and their genomic features used in the analysis are listed in [Table 5](#). A gene-level CN was calculated for each gene and each sample using TCGA-Assembler. The methylation value of a gene is the average methylation level of CpG sites that are DNase hypersensitive and are within 1500 base-pairs upstream of the transcription start site of the gene.

2.4 Statistical Methods

All genomic and radiomic imaging features were standardized to have mean 0 and standard deviation 1 prior to the subsequent analyses. Two types of statistical tests were conducted on the radiogenomic data.

First, *t* test was employed to investigate the differences of mean values in the different subgroups of patients as defined by the clinical outcomes. The Benjamini–Hochberg procedure³³ was used to control the false discovery rate (FDR) for the tests of each clinical outcome with all radiomic features or all genomic features. Adjusted *p*-values no larger than 0.1 were considered significant.

Second, logistic regression was used to model the relationship between clinical outcomes and radiogenomic features. The number of regressors (indicated by P) is relatively large, i.e., $P = 38$ for radiomic imaging data and $P = 144$ for genomic data. Logistic regression was conducted with LASSO regularization¹³ as the variable selection method to identify the features that best predict clinical outcomes. The LASSO method is a shrinkage and variable selection method for regression models. It maximizes a penalized log-likelihood function, which can be transformed into the following optimization problem given a positive value of λ :

$$\operatorname{argmax}_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^n \{ y_i \cdot (\beta_0 + \beta^T x_i) - \ln[1 + \exp(\beta_0 + \beta^T x_i)] \} - \lambda \|\beta\|_1 \right\},$$

where y_i is a $\{0, 1\}$ indicator for the clinical phenotype of patient i , β_0 is the intercept in the logistic model, β is the coefficient vector in logistic regression, x_i denotes the radiomic

profile or genomic profile of patient i , and λ is the tuning parameter determining the number of nonzero coefficients. After optimization, only salient features contributing to the discrimination between different clinical phenotypes will have nonzero β coefficients. Genomic features and radiomic features were investigated separately and combined in the logistic regression with LASSO regularization in order to select the best genomic predictors, the best radiomic predictors, and the best predictors among all radiogenomic features. In addition, the area under the receiver operating characteristic (ROC) curve (AUC) was obtained under cross-validation and is reported as the performance metric for prediction accuracy.

Because LASSO requires tuning of the model parameter λ , which controls the strength of regularization, a two-tier cross-validation was implemented to ensure the high quality of model training and to evaluate the generalization prediction performance, as illustrated in [Fig. 2](#). The inner-tier cross-validation was used to select the best λ value with the highest AUC on the testing data in the inner-tier cross-validation, and the outer-tier cross-validation measured the generalization performance of the prediction scheme. For each clinical outcome, the same number of data folds was used for both inner-tier cross-validation and outer-tier cross-validation. Since some of the clinical phenotypes were quite unbalanced (e.g., 77 ER+ versus 14 ER–), when splitting the data into training and testing sets, the percentage of samples with a given phenotype was kept the same in both training and testing sets as in the original whole dataset. [Figure 2](#) gives the flow chart showing the details of the

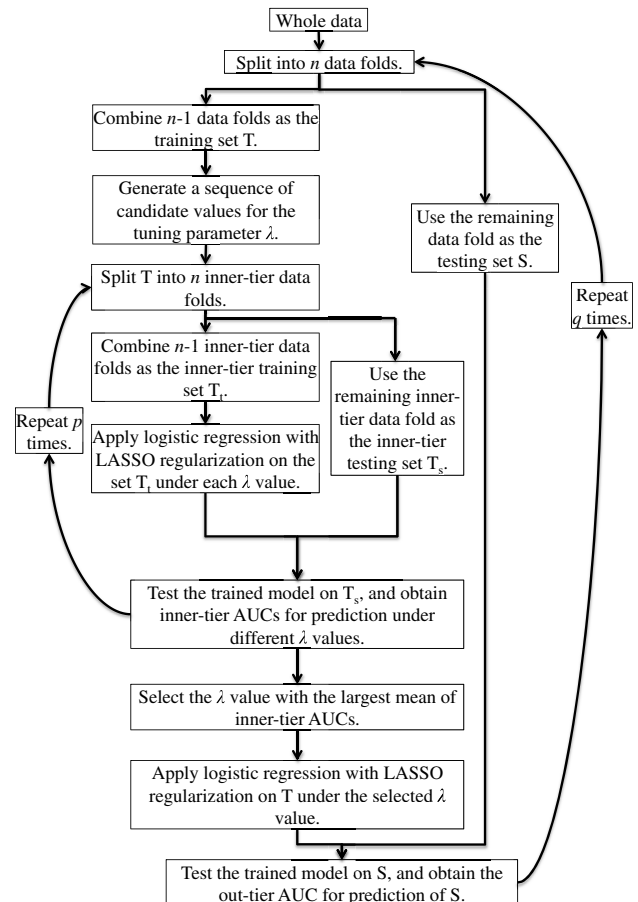


Fig. 2 Flow chart of the two-tier cross-validation.

Table 2 The numbers of data folds and cross-validation trials used in the two-tier cross-validation.

Clinical outcome	Number of data folds	Number of out-tier cross-validation trials	Number of inner-tier cross-validation trials
Stage	5	$5 \times 20 = 100$	$100 \times (5 \times 20) = 10,000$
Lymph node metastasis	10	$10 \times 10 = 100$	$100 \times (10 \times 10) = 10,000$
Estrogen receptor (ER)	5	$5 \times 20 = 100$	$100 \times (5 \times 20) = 10,000$
Progesterone receptor (PR)	8	$8 \times 12 = 96$	$96 \times (8 \times 12) = 9216$
HER2	5	$5 \times 20 = 100$	$100 \times (5 \times 20) = 10,000$

two-tier cross-validation scheme. The numbers of data folds and total cross-validation trials used for each clinical outcome are listed in Table 2.

We consider the AUC of each cross-validation trial as a sample from the AUC distribution and present the mean (mean_{AUC}) and standard deviation (sd_{AUC}) of sampled AUCs in Table 3.

We used R to carry out the analysis. R package “glmnet” was used for LASSO analysis and R package “pROC” was used for ROC analysis. We have provided a formal workflow tool for the radiogenomic analysis in our paper.³⁴

3 Results

Figure 3 shows the radiomic features whose mean values significantly (adjusted p -value ≤ 0.1) changed between clinical types as evaluated by t tests. Statistically significant associations are represented by edges. If ER+, PR+, HER2+, stage III, and positive lymph node metastasis are defined as the higher classes for the corresponding clinical outcome, a red edge means that the radiomic feature has a significantly larger mean value in the higher class than in the lower class, while a blue edge means the opposite. Note that all tumor size features (S category in Fig. 3) are significantly positively associated with tumor stage, showing that tumor size is one of the major factors considered in the current tumor staging system. Tumor shape feature G2 (irregularity) is significantly positively associated with tumor stage, indicating that higher-stage tumors have more irregular shape. One tumor margin feature M3 (variance of radial gradient histogram) and two enhancement texture features T7 (inverse difference moment) and T11 (sum average) are also predictive of tumor stage. Enhancement texture T5 (angular second moment—energy) is predictive of PR status. No radiomic feature is significantly associated with other clinical outcomes, including lymph node metastasis, ER, and HER2 status. Please check Table 1 for the category label and index of radiomic features.

Using t tests, we also identify the genomic features with significantly differential mean values between high and low classes of clinical outcomes (see Table 4). There is no genomic feature that significantly (adjusted p -value ≤ 0.1) differentiates phenotypes of tumor stage and lymph node metastasis. A lot of genomic features are significantly associated with ER and PR status, and only one genomic feature, TP53(ME), is significantly associated with HER2 status.

Results of LASSO-based logistic regression analysis are given in Table 3. Genomic features performed better than radiomic features in predicting ER and PR status, with average AUCs of 0.916 and 0.775 from cross-validations, respectively, while radiomic features performed better in predicting pathological

stage with an average AUC of 0.877. These results were expected as ER and PR statuses are genomic types closely related to tumor genomic profiles while pathological stage is clinically defined on phenotypes of the tumors, some of which can be directly characterized by imaging. For prediction of lymph node metastasis and HER2 status, neither genomic features nor radiomic imaging features did well (with average AUCs ≤ 0.7). The most discriminative radiomic feature to predict less-aggressive stage I tumors versus more-aggressive stage III tumors was effective diameter, which measures the tumor size. Larger values of effective diameter are usually an important sign of more-aggressive tumors.³⁵ Our results agree with this observation, as the coefficient of effective diameter in the logistic prediction model, which is trained based on effective diameter alone and all tumor samples for predicting tumor pathological stage, is positive. The most discriminative genomic feature to predict tumor ER status is AURKB(GE), and its coefficient in the logistic model trained based on AURKB(GE) alone and all tumor samples for predicting ER status is negative.

Comparison of AUCs obtained on the integrated radiogenomics data with those obtained on genomic data or radiomic data alone indicated no improvement in the prediction accuracy by combining two different data modalities. For the case of pathological stage, ER, and PR status, the reason could be that no feature from the less-predictive data modality can provide a complementary prediction power to the most-predictive features from the more-predictive data modality. Both the most frequently selected feature set and the most frequently selected individual features do not change or change very little between the more-predictive data modality and the integrated data. For lymph node metastasis and HER2 status, it seems that both genomic data and radiomic data lack the power for a good prediction and their integration did not show any improvement.

4 Discussion

A comprehensive analysis was conducted on the integration of genomic and radiomic data of 91 breast cancer patients from TCGA and TCIA. We believe that our study is the largest study to date that combines multiple types of genomic data with radiomic data in predicting breast cancer prognosis. Relationships were explored between the genomic and radiomic features and five selected clinical outcomes categorized the tumors into different subgroups related to prognosis and treatment scheme.

The single variable t test identified that all tumor size features are significantly associated with the tumor pathological stage showing the importance of size in the stage classification of tumors. However, no individual radiomic feature was found

Table 3 Prediction and feature selection results on genomic data and radiomic data separately and combined. For each prediction analysis, the median number of selected features, the mean and standard deviation of areas under the receiver operating characteristic curve (AUCs), the most frequently selected feature set, and the individual features selected in more than 50% of the cross-validation trials, all based on the outer-tier cross-validation, are reported.

Clinical outcome	Result category	Genomics	Radiomics	Genomics + Radiomics
Stage	Median number of selected features	7	2	1
	mean _{AUC} (sd _{AUC})	0.647 (0.155)	0.877 (0.161)	0.870 (0.185)
	Most frequent feature set and its frequency	epidermal growth factor receptor (EGFR)(CN), $f = 0.07$	Effective diameter, $f = 0.36$	Effective diameter, $f = 0.59$
	Individual features with frequency ≥ 0.5	EGFR(CN), PPP2R2B(GE), TBX3(GE), AFF2(GE), BCL2(GE)	Effective diameter	Effective diameter
Lymph node	Median number of selected features	44.5	4.5	32
	mean _{AUC} (sd _{AUC})	0.654 (0.131)	0.693 (0.156)	0.634 (0.139)
	Most frequent feature set and its frequency	TBX3(CN) PTPN22(GE), $f = 0.02$	Irregularity, inverse difference moment, and variance of radial gradient histogram, $f = 0.14$	Irregularity, $f = 0.02$; effective diameter, $f = 0.02$
	Individual features with frequency ≥ 0.5	PAK1(GE), TP53(ME), MYC(GE), PIK3CA(CN), CCND3(GE), PTPN22(GE), AFF2(ME), PTEN(ME), MTHFD1L(GE), CTCF(GE), CCND3(CN), ZNF703(GE), PPP2R1A(GE), PTEN(GE), PPP2R2A(GE), CHEK1(CN), ERBB2(CN), MTAP(CN), PPP2R2B(GE), GATA3(CN), RB1(GE)	Irregularity, inverse difference moment, variance of radial gradient histogram, variance of margin sharpness	Irregularity, MYC(GE), inverse difference moment, PAK1(GE), AFF2(ME), ERBB2(CN), CTCF(GE), sum average, CCND3(GE), TP53(ME), variance of margin sharpness, PPP2R2B(GE), CCND3(GE), MTHFD1L(GE), PTEN(ME), CCND3(CN), PIK3CA(CN), GATA3(GE), ERBB2(GE), PTPN22(GE)

Table 3 (Continued).

Clinical outcome	Result category	Genomics	Radiomics	Genomics + Radiomics
ER	Median number of selected features	2	5	2
	mean _{AUC} (sd _{AUC})	0.916 (0.095)	0.789 (0.140)	0.915 (0.101)
	Most frequent feature set and its frequency	AURKB(GE), $f = 0.27$	Effective diameter, angular second moment (energy), sphericity, difference variance, time to peak at maximum variance, sum average, $f = 0.09$	AURKB(GE), $f = 0.26$
	Individual features with frequency ≥ 0.5	AURKB(GE)	Effective diameter, angular second moment (energy), sphericity, difference variance, time to peak at maximum variance, sum average	AURKB(GE)
PR	Median number of selected features	25.5	10	28
	mean _{AUC} (sd _{AUC})	0.775 (0.164)	0.689 (0.160)	0.760 (0.171)
	Most frequent feature set and its frequency	AURKB(GE), $f = 0.06$	Angular second moment (energy), $f = 0.09$	AURKB(GE), $f = 0.04$
	Individual features with frequency ≥ 0.5	CDK2(GE), CAMK1D(CN), CDK4(CN), ZNF703(GE), TBX3(CN), CBFB(GE), BRIP1(CN), ERBB2(GE), ATM(ME), CDC45(CN), PPP2R2A(CN), AFF2(CN), AURKB(GE), RB1(GE), CCND3(GE)	Angular second moment (energy), lesion volume, effective diameter, time to peak at maximum variance, margin sharpness, time to peak, sphericity, volume of most enhancing voxels, maximum linear size	CDK2(GE), CAMK1D(CN), ZNF703(GE), angular second moment (energy), CDK4(CN), CBFB(GE), ERBB2(GE), KIFC1(GE), CCND3(CN), RB1(GE), ATM(ME), TBX3(CN), BRIP1(CN), surface-to-volume ratio, CDC45(CN), HDAC2(GE), AFF2(CN), AURKB(GE), AFF2(ME), CCNE1(GE)
HER2	Median number of selected features	4	3	5
	mean _{AUC} (sd _{AUC})	0.635 (0.135)	0.641 (0.125)	0.611 (0.143)
	Most frequent feature set and its frequency	NBN(GE), $f = 0.09$	Variance of radial gradient histogram, $f = 0.11$	NBN(GE), $f = 0.09$
	Individual features with frequency ≥ 0.5	NBN(GE), TP53(CN), AKT1(CN), CBFB(CN)	Variance of radial gradient histogram, time to peak	NBN(GE), TP53(CN), CBFB(CN), AKT1(CN)

Note: CN, copy number; GE, gene expression; ME, DNA methylation.

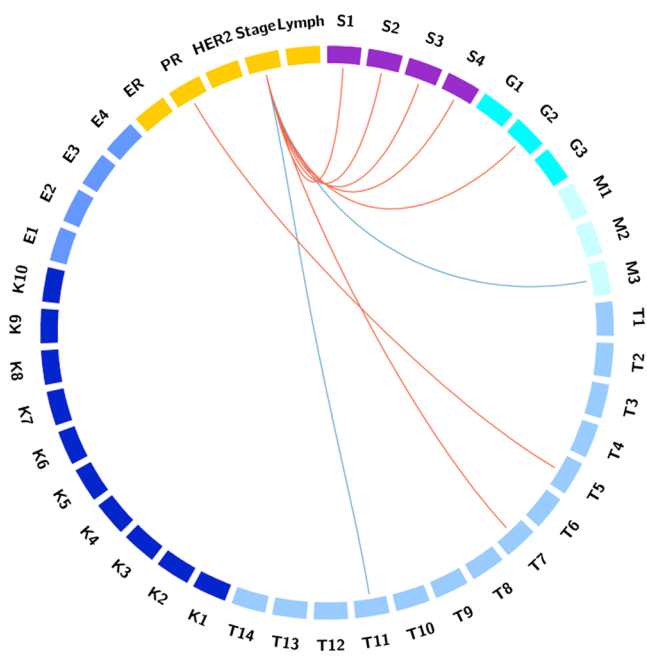


Fig. 3 Significant associations between radiomic features and clinical outcomes evaluated by *t*-tests. Names and descriptions of radiomic features can be found in Table 1. Red edges indicate higher feature values in the higher class of the clinical outcome, and blue edges indicate the opposite. The higher classes are ER+, PR+, HER2+, stage III, and positive lymph node metastasis.

to be significantly associated with lymph node metastasis, ER, and HER2 status. On the other hand, no individual genomic feature showed significant association with tumor stage and lymph node metastasis. We found many more genomic features significantly associated with ER and PR status than with HER2 status.

Using logistic regression with LASSO regularization, the effective diameter was selected as the most-predictive feature for tumor pathological stage, outperforming the genomic features in this clinical task as expected given the connection between the current tumor staging classification and tumor size. Genomic feature AURKB(GE) was selected as the most-predictive feature of ER status. Genomic features outperformed radiomic features in predicting both ER status and PR status, both of which are molecular characteristics.

We observed that radiomic features were more predictive for tumor stage than lymph node involvement. Note that the current tumor staging system is based on the T-N-M status of the tumor. Tumor T status describes the size of the original (primary) tumor and whether the tumor has invaded nearby tissue, which has an important role in tumor staging and can be characterized by MRI. It is not surprising to see that radiomic features have a high AUC for predicting tumor stage due to the correlation between tumor T status and tumor stage. However, lymph node involvement, i.e., the tumor N status, may not be easily characterized by MRI of the primary tumor, since it does not account for or analyze the number of positive lymph nodes. Thus, radiomic features cannot provide a good prediction of lymph node involvement.

Overall, the prediction performance by genomic features alone, radiomic features alone, and combined radiogenomics features showed significant correlations with clinical outcomes. However, the change in predictive performance when going from either genomic features alone or radiomic features alone to the combined radiogenomic features was not found to be statistically significant, most likely due to the limited data size (91 cancer cases with 38 radiomic features and 144 genomic features) and the types of clinical variables that we considered. Tumor stage is a phenotypic variable closely related to tumor

Table 4 Genomic features significantly associated with clinical outcomes. “+” indicates larger feature values in the higher class, while “-” indicates the opposite.

Clinical outcome	Genomic features
Stage	None
Lymph node metastasis	None
ER	<p>+: MDM1(CN), CDK4(CN), PPP2R2B(CN), MAP2K4(GE), BCL2(GE), ZNF703(GE), PTEN(CN), PTEN(GE), CCND1(GE), MDM2(CN), MDM2(GE), TP53(ME), GATA3(GE), RB1(GE), MAP3K1(CN), MAP3K1(GE), TBX3(GE), TBX5(GE), FOXA1(GE), RUNX1(GE), PIK3R1(CN), PTPRD(GE), NF1(GE), BRIP1(CN).</p> <p>–: CDK3(GE), CAMK1D(CN), MTAP(GE), CDKN2A(GE), AURKB(GE), BUB1(GE), CDCA3(GE), CDCA4(GE), CDC20(CN), CDC20(GE), CDC45(GE), CHEK1(GE), FOXM1(GE), HDAC2(CN), HDAC2(GE), KIF2C(CN), KIF2C(GE), KIFC1(GE), MTHFD1L(CN), MTHFD1L(GE), RAD51AP1(CN), RAD51AP1(GE), TTK(CN), TTK(GE), UBE2C(GE), CCNE1(GE), GATA3(CN), CDKN1B(CN), CTCF(CN), CBFB(CN), CBFB(GE), CHEK2(GE)</p>
PR	<p>+: MDM1(CN), MDM1(GE), CDK4(CN), PPP2R2B(CN), BCL2(GE), PTEN(GE), MDM2(CN), MDM2(GE), GATA3(GE), RB1(GE), MAP3K1(CN), MAP3K1(GE), TBX3(CN), TBX3(GE), TBX5(CN), FOXA1(GE), RUNX1(GE), AFF2(ME), PIK3R1(CN), PTPRD(GE), NF1(GE).</p> <p>–: CDK3(GE), CAMK1D(CN), AURKB(GE), BUB1(GE), CDCA3(GE), CDCA4(GE), CDC20(GE), CDC45(CN), CDC45(GE), CHEK1(GE), FOXM1(GE), HDAC2(GE), KIF2C(GE), KIFC1(GE), MTHFD1L(GE), TTK(GE), UBE2C(GE), CCNE1(GE), GATA3(CN), CBFB(GE), CHEK2(CN), CHEK2(GE)</p>
HER2	–: TP53(ME)

size and invasion characterized by MRI. Thus, radiomic features alone already provided a good prediction. Compared to radiomic features, genomic features may have a weaker correlation with tumor stage and, thus, did not add additional prediction power. On the other hand, for clinical variables related to the genomic status of a tumor, such as ER status and PR status, it was not surprising to find that genomic features have more predictive power, while radiomic features do not provide additional predictive power. However, for other types of clinical variables, such as survival, there is the potential for combined genomic and radiomic features to provide a better prediction than each type of feature alone, although, unfortunately, on our limited data with only one terminal event, we could not assess this.

In the future, we plan to collect more tumor samples to study in depth whether combining radiomic and genomic features would improve the prediction of clinical profiles. Currently, the clinical outcomes are taken as binary variables, but some of them actually have multiple outcome values, for example, the tumor pathological stage. In the future work, we will use multinomial regression for the analysis of these clinical variables. Also, we will consider the dependence among features when doing feature selection and prediction to coincide with the collinearity among features.

Balancing techniques are useful when data are imbalanced and the precision and recall are of different importance. In our analysis, since the ROC curve (and hence AUC) is insensitive to changes in class imbalance,²¹ we did not make use of balancing techniques in the analysis. Besides, if balancing techniques were utilized, we would have to reinterpret the AUC because of the trade-off between the precision and recall induced by using balancing techniques, such as SMOTE.³⁶ These can be considered in future work if a trade-off between the precision and recall is necessary.

Radiogenomics is an emerging new field for cancer research. Our results serve as an initial attempt in the radiogenomics of breast cancer and provide guidance for future investigations. We did not investigate the relationship between radiogenomic features and patient survival since only one mortality event existed among the 91 patients. The power of the presented analysis is bounded by the small sample size of 91 patients. As the community starts to accumulate more data, larger studies are expected to shed more light on the relationship between radiogenomic features and clinical outcomes.

Appendix: Information About Genomic Features

1. Gene expression (GE) is the process in which the genetic information (DNA code) of a gene is transcribed into a messenger RNA (mRNA), which further serves as a template used in the synthesis of a functional gene product. Usually, the functional gene product is a protein for protein coding genes. GE in our analysis refers to the level of gene expression, which is the quantity of mRNAs that is transcribed from a gene.
2. Copy number (CN) in our analysis refers to the number of copies of a gene in the genome. In cancer, the CN of a gene may change, which is called CN variation. It is a form of genetic structural variation of DNA

Table 5 The 70 genes and their features used in the analysis. CN, GE, and ME stand for gene copy number, gene expression, and DNA methylation, respectively.

Gene name	Platform	Gene name	Platform
MDM1	CN, GE	MYC	CN, GE
MDM4	CN, GE	CCND1	CN, GE
CDK3	CN, GE	MDM2	CN, GE
CDK4	CN, GE	ERBB2	CN, GE
CAMK1D	CN, GE	CCNE1	CN, GE
PI4KB	CN, GE	PIK3CA	CN, GE
NCOR1	CN, GE	AKT1	CN, GE
PPP2R1A	CN, GE	TP53	CN, GE, ME
PPP2R2A	CN, GE	GATA3	CN, GE
PPP2R2B	CN, GE	CDH1	CN, GE
MTAP	CN, GE	RB1	CN, GE
CDKN2A	CN, GE	MAP3K1	CN, GE
CDKN2B	CN, GE	CDKN1B	CN, GE
MAP2K4	CN, GE	TBX3	CN, GE
PAK1	CN, GE	TBX4	CN, GE
RSF1	CN, GE	TBX5	CN, GE
AURKB	CN, GE	CTCF	CN, GE
BCL2	CN, GE	FOXA1	CN, GE
BUB1	CN, GE	RUNX1	CN, GE
CDCA3	CN, GE	CBFB	CN, GE
CDCA4	CN, GE	AFF2	CN, GE, ME
CDC20	CN, GE	PIK3R1	CN, GE
CDC45	CN, GE	PTPN22	CN, GE
CHEK1	CN, GE	PTPRD	CN, GE
FOXM1	CN, GE	NF1	CN, GE
HDAC2	CN, GE	SF3B1	CN, GE
IGF1R	CN, GE	CCND3	CN, GE
KIF2C	CN, GE	ATM	CN, GE, ME
KIFC1	CN, GE	BRCA1	CN, GE
MTHFD1L	CN, GE	BRCA2	CN, GE
RAD51AP1	CN, GE	BRIP1	CN, GE
TTK	CN, GE	CHEK2	CN, GE
UBE2C	CN, GE	NBN	CN, GE
ZNF703	CN, GE	RAD51C	CN, GE
PTEN	CN, GE, ME	EGFR	CN, GE

that results in the change of the number of copies of a gene's DNA segment.

3. DNA methylation (ME) is a biochemical process where a methyl group is added to the cytosine or adenine DNA nucleotides. In adult somatic cells (cells in the body not used for reproduction), ME typically occurs in a CpG dinucleotide context. ME can lead to various effects, such as inhibiting the transcription of genes. In our study, the ME of a gene is the average methylation level of CpG sites that are DNase hypersensitive and are within 1500 base-pairs upstream of the transcription start site of the gene (see Table 5).

Acknowledgments

Members of the TCGA Breast Phenotype Research group also include Elizabeth Morris, Elizabeth Burnside, Emelinda Bonaccio, Marie Ganott, Jose Net, Elizabeth Sutton, Gary Whitman, Margarita Zuley, Kathy Brandt, Carl Jaffe, Erich Huang, John Freymann, and Justin Kirby. Wentian Guo is currently a graduate student at Fudan University, School of Public Health, Shanghai, China. Maryellen L. Giger's research was partly supported by a University of Chicago Dean Bridge Fund. She is a stockholder in R2 Technology/Hologic and receives royalties from Hologic, GE Medical Systems, MEDIAN Technologies, Riverain Medical, Mitsubishi, and Toshiba. She is a cofounder of and stockholder in Quantitative Insights. Yuan Ji's research is partly supported by NIH 2R01 CA132897.

References

1. H. J. Aerts et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nat. Commun.* **5**, 4006 (2014).
2. B. S. Rosenstein et al., "Radiogenomics: radiobiology enters the era of big data and team science," *Int. J. Radiat. Oncol. Biol. Phys.* **89**(4), 709–713 (2014).
3. B. F. Kurland et al., "Promise and pitfalls of quantitative imaging in oncology clinical trials," *Magn. Reson. Imaging* **30**(9), 1301–1312 (2012).
4. P. Lambin et al., "Predicting outcomes in radiation oncology—multifactorial decision support systems," *Nat. Rev. Clin. Oncol.* **10**(1), 27–40 (2013).
5. L. H. Sobin, M. K. Gospodarowicz, and C. Wittekind, Eds., "TNM Classification of Malignant Tumours," John Wiley & Sons, Hoboken, New Jersey (2011).
6. R. Rouzier et al., "Breast cancer molecular subtypes respond differently to preoperative chemotherapy," *Clin. Cancer Res.* **11**(16), 5678–5685 (2005).
7. F. J. Esteva et al., "Molecular predictors of response to trastuzumab and lapatinib in breast cancer," *Nat. Rev. Clin. Oncol.* **7**(2), 98–107 (2010).
8. O. Gevaert et al., "Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results," *Radiology* **264**(2), 387–396 (2012).
9. N. Jamshidi et al., "Illuminating radiogenomic characteristics of glioblastoma multiforme through integration of MR imaging, messenger RNA expression, and DNA copy number variation," *Radiology* **270**(1), 1–2 (2014).
10. C. A. Karlo et al., "Radiogenomics of clear cell renal cell carcinoma: associations between CT imaging features and mutations," *Radiology* **270**(2), 464–471 (2014).
11. J. N. Weinstein et al., "The Cancer Genome Atlas Pan-Cancer Analysis project," *Nat. Genet.* **45**(10), 1113–1120 (2013).
12. K. Clark et al., "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *J. Digit. Imaging* **26**(6), 1045–1057 (2013).
13. T. T. Wu et al., "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics* **25**(6), 714–721 (2009).
14. Y. Zhu, P. Qiu, and Y. Ji, "TCGA-Assembler: open-source software for retrieving and processing TCGA data," *Nat. Methods* **11**(6), 599–600 (2014).
15. 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).
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. A. Shimauchi et al., "Evaluation of clinical breast MR imaging performed with prototype computer-aided diagnosis breast MR imaging workstation: reader study," *Radiol.* **258**(3), 696–704 (2011).
19. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. of the 14th Int. Joint Conf. on Artificial Intelligence*, pp. 1137–1143 (1995).
20. S. Borra and D. C. Agostino, "Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods," *Comput. Stat. Data Anal.* **54**(12), 2976–2989 (2010).
21. T. Fawcett, "ROC graphs: notes and practical considerations for researchers," Technical Report HPL-2003-4, Hewlett Packard Labs (2004).
22. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
23. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.* **3**, 610–621 (1973).
24. W. Chen et al., "Computerized interpretation of breast MRI: investigation of enhancement-variance dynamics," *Med. Phys.* **31**, 1076–1082 (2004).
25. 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**, 822–829 (2010).
26. N. Bhooshan et al., "Cancerous breast lesions on dynamic contrast-enhanced MR images: computerized characterization for image-based prognostic markers," *Radiology* **254**(3), 680–690 (2010).
27. N. Bhooshan et al., "Computerized three-class classification of MRI-based prognostic markers for breast cancer," *Phys. Med. Biol.* **45**, 5995–6008 (2011).
28. C. Curtis et al., "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature* **486**(7403), 346–352 (2012).
29. Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature* **490**(7418), 61–70 (2012).
30. K. Wang et al., "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic Acids Res.* **38**, e178 (2010).
31. B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome," *BMC Bioinf.* **12**, 323 (2011).
32. A. B. Olshen, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics* **5**, 557–572 (2004).
33. K. G. A. Gilhuijs, M. L. Giger, and U. Bick, "Automated analysis of breast lesions in three dimensions using dynamic magnetic resonance imaging," *Med. Phys.* **25**, 1647–1654 (1998).
34. http://compgenome.org/TCGA/R_code_package.zip
35. C. L. Carter, C. Allen, and D. E. Henson, "Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases," *Cancer* **63**(1), 181–187 (1989).
36. N. V. Chawla et al., "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.* **16**, 321–357 (2002).

Wentian Guo is a PhD candidate in the School of Public Health, Fudan University, Shanghai, China. She is currently a visiting student at the University of Chicago. Her research interests include statistical modeling, clinical trial design, and bioinformatics.

Hui Li has been working on quantitative imaging analysis at the University of Chicago since 2001. His research interests include breast cancer risk assessment and computer-aided diagnosis on

mammography and magnetic resonance imaging, understanding the relationship between image-based phenotypes and genomics and their future roles in personalized medicine.

Yitan Zhu is a research scientist at NorthShore University HealthSystem, Illinois. His research interests include bioinformatics, systems biology, drug development, machine learning, signal processing, and image analysis. Currently, his major research projects focus on processing and analyzing The Cancer Genome Atlas (TCGA) data.

Li Lan has been working on breast image analysis research at the University of Chicago since 1996. Her research interests include developing user-friendly workstations/software packages, database management, and data analysis.

Shengjie Yang is currently a postdoc fellow at NorthShore University HealthSystem, Illinois. He obtained his PhD in bioinformatics from Fudan University, China. His research focuses on developing functionality web tool and innovative statistical designs for clinical trials in oncology, and building pipelines for analysis of cancer genomics using TCGA data.

Karen Drukker has been involved in medical image analysis research for over a decade, with a focus on quantitative imaging in the detection, diagnosis, and prognosis of breast cancer. Apart from coauthoring many publications in respected peer-reviewed journals, she has coauthored three medical imaging book chapters and is listed on two U.S. patents.

Elizabeth Morris is chief of the breast imaging service at Memorial Sloan Kettering Cancer Center (MSKCC) and professor of radiology at the Weill Medical College of Cornell University. Her research interests focus on the MRI detection of early-stage breast cancer in high-risk women and in MRI-guided procedures. She authored the book *Breast MRI: Diagnosis and Intervention*. Her recent research efforts have involved looking at imaging biomarkers to assess risk and treatment response.

Elizabeth Burnside is currently a professor of radiology in the University of Wisconsin School of Medicine and Public Health. Her MD degree combined with master's degrees in public health and medical informatics provide a foundation for her research that investigates the use of machine learning and artificial intelligence methods to improve decision-making in the domain of breast imaging in the pursuit of improving the population-based screening and diagnosis of breast cancer.

Gary Whitman is professor of radiology and radiation oncology at the University of Texas MD Anderson Cancer Center (MDACC), where he serves as the medical director of the mobile mammography program. He joined the MDACC faculty in 1996 after serving on the faculty at Harvard Medical School and Massachusetts General Hospital. He is the past chair of the MDACC faculty senate, and he serves on the MDACC shared Governance Committee.

Maryellen L. Giger is the A. N. Pritzker professor of radiology and is on the Committee on Medical Physics at the University of Chicago. Her research interests mainly involve the investigation of computer-aided diagnosis and radiomic methods for the assessment of risk, diagnosis, prognosis, and response to therapy of breast cancer on multimodality (mammography, ultrasound, and magnetic resonance) images. She is also involved in broad-based developments in computer vision and data mining of medical images.

Yuan Ji is assistant vice president, director of computational genomics and medicine at NorthShore University HealthSystem. He is also associate professor (biostatistics, part-time) at the University of Chicago and adjunct associate professor of biostatistics at the University of Texas. His research focuses on innovative computational and statistical methods for translational cancer research. He has authored nearly 100 publications in peer-reviewed journals, conference papers, book chapters, and abstracts.