Eigentumors for prediction of treatment failure in patients with early-stage breast cancer using dynamic contrast-enhanced MRI

A feasibility study

H. M. Chan
Image Sciences Institute, University Medical Center Utrecht, Q.02.4.45, P.O. Box 85500, 3508 GA Utrecht, The Netherlands

B. H. M. van der Velden
Image Sciences Institute, University Medical Center Utrecht, Q.02.4.45, P.O. Box 85500, 3508 GA Utrecht, The Netherlands

C. E. Loo
Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, P. O. Box 90203, 1006 BE Amsterdam, The Netherlands

K. G. A. Gilhuijs
Image Sciences Institute, University Medical Center Utrecht, Q.02.4.45, P.O. Box 85500, 3508 GA Utrecht, The Netherlands

Abstract

We present a radiomics model to discriminate between patients at low risk and those at high risk of treatment failure at long-term follow-up based on eigentumors: principal components computed from volumes encompassing tumors in washin and washout images of pre-treatment dynamic contrast-enhanced (DCE-) MR images. Eigentumors were computed from the images of 563 patients from the MARGINS study. Subsequently, a least absolute shrinkage selection operator (LASSO) selected candidates from the components that contained 90% of the variance of the data. The model for prediction of survival after treatment (median follow-up time 86 months) was based on logistic regression. Receiver operating characteristic (ROC) analysis was applied and area-under-the-curve (AUC) values were computed as measures of training and cross-validated performances. The discriminating potential of the model was confirmed using Kaplan-Meier survival curves and log-rank tests.

From the 322 principal components that explained 90% of the variance of the data, the LASSO selected 28 components. The ROC curves of the model yielded AUC values of 0.88, 0.77 and 0.73, for the training, leave-one-out cross-validated and bootstrapped performances, respectively. The bootstrapped Kaplan-Meier survival curves confirmed significant separation for all tumors ($P < 0.0001$). Survival analysis on immunohistochemical subgroups shows significant separation for the estrogen-receptor (ER) subtype tumors ($P < 0.0001$) and the triple-negative (TN) subtype tumors ($P=0.0039$), but not for tumors of the HER2 subtype ($P=0.41$). The results of this retrospective study show the potential of early-stage pre-treatment eigentumors for use in prediction of treatment failure of breast cancer.

Introduction

Breast cancer is recognized as a highly heterogeneous disease, for which it is desirable to distinguish between more indolent cancer types and those that lead to poor patient survival. Characterization of breast-cancer type based on imaging may enable more effective treatment tailored to individual patients, thus reducing both undertreatment and overtreatment. Concern of overtreatment – treatment without survival benefit – exists especially in patients with early breast cancer eligible for breast-conserving surgery, where the question arises whether chemotherapy that may lead to adverse side effects should be given.

Currently, long-term survival still cannot be predicted well on individual patient basis. Tissue markers such as histologic tumor grade, mitotic activity index (MAI), and lymph node status are typically used to establish prognosis of the patient. In addition, three different breast cancer subtypes have been recognized based on the status of the cancer's hormone receptors, resulting in markedly different patient prognosis (Chen et al., 2010, Cianfrocca and Goldstein, 2004, de Mascarel et al., 2015, Sørlie et al., 2001). However, tissue markers may be prone to intra-tumoral heterogeneity if the tissue is sampled by core needle biopsy (Schmitz et al., 2014, Focke et al., 2016, Richter-Ehrenstein et al., 2009).

Imaging takes the entire in-situ tumor into consideration. Since visual assessment of MR images by human readers may suffer from inter-observer variability due to the subjective nature of image interpretation (de Camargo Moraes et al., 2010, Stoutjesdijk et al., 2005, Wedegartner et al., 2001), automated computerized image analysis may be preferable.

There have been studies investigating associations between imaging features and pathological complete response (pCR) to neoadjuvant chemotherapy as surrogate endpoint for survival. However, multiple studies have indicated that pCR is not always an accurate surrogate endpoint for survival (Hamy-Petit et al., 2016, Pennisi et al., 2016, von Minckwitz and Fontanella, 2015). Furthermore, neoadjuvant chemotherapy is typically given to patients with locally-advanced breast cancer; hence, imaging findings may not be directly comparable to those in an early-stage breast cancer population. In addition, reported associations between MR imaging and breast cancer patient outcome typically involve relatively small subgroups from heterogeneous or not clearly described populations.

In an approach to find relevant imaging features, we apply principal component analysis (PCA) to the full spatial domain in and around the tumor during contrast uptake to yield components that capture both the shape and enhancement characteristics. We call these eigenvectors of tumor images "eigentumors", analogous to eigenfaces used for facial recognition (Turk and Pentland, 1991), and hypothesize that a set of these eigentumors is correlated with long-term survival of the patient. To determine which eigentumors are relevant, we train directly on survival as endpoint. To the best of our knowledge, this approach has not been investigated for the analysis of breast tumors, neither for classification nor for prognostication.

This study has two aims: first, to identify a set of eigentumors in DCE-MRI using a large series of consecutively included patients with early breast cancer eligible for breast conserving therapy. The second aim is to construct a model to automatically discriminate between patients at low risk and those at high risk of treatment failure based on the identified eigentumors, and to assess the model's performance.

Methods

Patient cohort and follow-up

Our retrospective study is based on data from the MARGINS study (Multi-modality Analysis and Radiological Guidance IN breast conServing therapy) which was performed at the Netherlands Cancer Institute from 2000 to 2008. The patient population consisted of consecutively included women with pathology-proven early-stage breast cancer on pre-operative assessment, who were eligible for breast-conserving therapy. The patients were treated according to the Dutch national guidelines (www.oncoline.nl). Survival analysis was performed for overall survival (OS) using the definition that events include both deaths from cancer and from unknown causes (Hudis et al., 2007).

Pathology factors and imaging features

The patients' tumors were identified on the MR images, and for patients presenting multiple tumors, the largest lesion was used. The tumors were divided into three immunohistochemical (IHC) subtypes based on status of the estrogen receptor (ER), progesteron receptor (PR) and human epidermal growth factor receptor 2 (HER2). Tumors that are negative for all three receptors are of the triple-negative

(TN) subtype. Tumors that are ER-positive and HER2-negative belong to the ER-subtype, and HER2-subtype tumors are HER2-positive and ER-negative. ER and PR-status were determined on immunohistochemistry using hematoxylin and eosin stained microscopic sections, where a threshold of staining was used for the division of cases in negative (<10% staining) and positive (≥10% staining). HER2-amplification was scored as 0, 1+, 2+ or 3+. Cases with scores 0 and 1+ were classified as HER2-negative, and cases with score 3+ as HER2-positive. For cases with a 2+ score, fluorescent in-situ hybridization was used to determine the HER2-status.

The number of positive lymph nodes were determined by sentinel node biopsy, and combined with axillary lymph node dissection where available. The cases were grouped into three categories: no (0), one to three (1-3), or four or more (>4) positive lymph nodes. Histologic tumor grade was assessed with the modified Bloom-Richardson guidelines where morphology of the tubule and gland formation, nuclear pleomorphism and mitotic count are taken into account (Rakha et al., 2008).

Tumor-segmentation masks were obtained previously for the MARGINS data using the semi-automatic method of Alderliesten et al. (Alderliesten et al., 2007). Using these segmentations, four imaging features evaluating the tumor shape and tumor enhancement were computed: the circularity, the irregularity, the uptake speed and the washout, as described previously by Gilhuijs et al. (Gilhuijs et al., 2002, Gilhuijs et al., 1998). The fraction of tumor voxels in the original (non-scaled) region of interest was also calculated, where the fraction was defined as number of tumor voxels divided by the total number of voxels in the region of interest.

Significance of variation between subtypes was tested by Kruskal-Wallis and Fisher exact tests using version 0.17.0 of the SciPy module (Jones et al., 2001) in Python version 2.7.11 (Python Software Foundation, Beaverton, USA). *P*-values lower than 0.05 were considered significant.

Magnetic resonance imaging

The acquisition of breast DCE-MR images adhered to the guidelines for breast DCE-MR imaging by the European Society of Breast Imaging published in 2008 (Mann et al., 2008). The DCE-MR images were coronal fast low-angle shot three-dimensional T1-weighted images, acquired by a 1.5 T imaging unit (Magnetom; Siemens, Erlangen, Germany) with a dedicated double breast array coil (CP Breast

array, four channels; Siemens). A bolus of gadolinium-based contrast agent (Prohance; Bracco-Byk Gulden, Konstanz, Germany) was administered at 3 ml/s using a power injector. Five series of images were acquired: one unenhanced series before and four enhanced series after administration of contrast. The acquisition time per volume was 90 s. The repetition time was 8.1 ms and the echo time was 4.0 ms. A flip angle of 20° was used. The voxel size was 1.35 mm × 1.35 mm × 1.35 mm, and field of view was 310 mm.

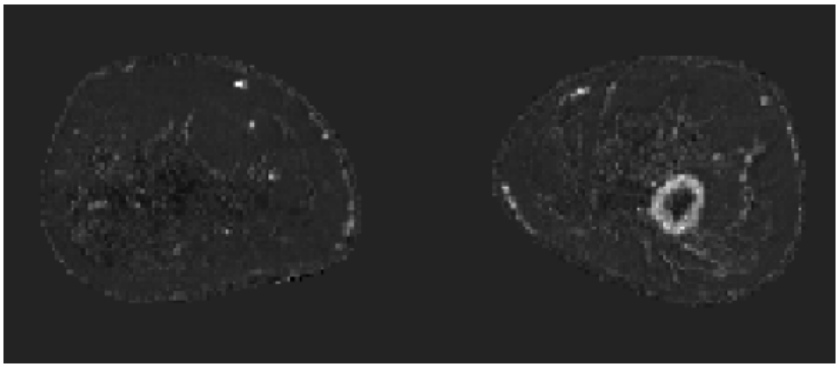
Construction of the predictive model

The steps for arriving at a model predictive of survival can be summarized as follows: first, the washin and washout images are computed (Figure 1). From the tumor areas, feature vectors are extracted (Figure 2). These feature vectors are used in training a model that predicts the probability of survival for a patient (Figure 3), a process which includes computing the principal components of the data and applying the least absolute shrinkage selection operator (LASSO) (Tibshirani, 1996). The following paragraphs will provide further details about our method. For the construction of the model, version 0.17.1 of the scikit-learn module (Pedregosa et al., 2011) in Python version 2.7.11 (Python Software Foundation, Beaverton, USA) was used.

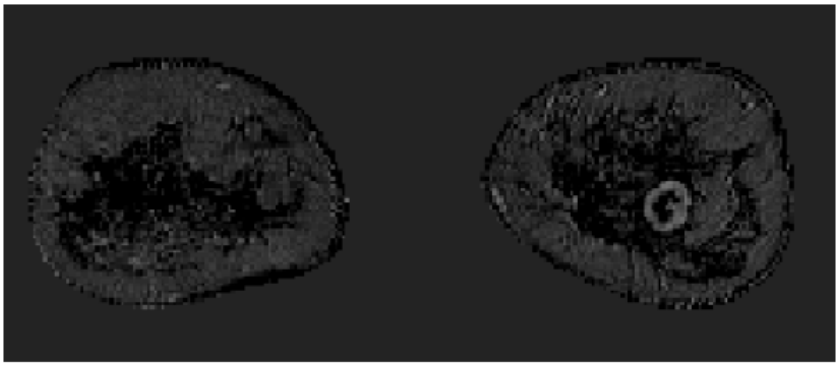
Washin and washout images I_{washin} and I_{washout} (Figure 1) were computed from the dynamic contrast series for each case using the following formulas

$$I_{washin}(k) = \frac{I_{early}(k) - I_{pre}(k)}{I_{pre}(k)} \cdot 100\% \text{ and } \qquad I_{washout}(k) = \frac{I_{early}(k) - I_{late}(k)}{I_{early}(k)} \cdot 100\%,$$

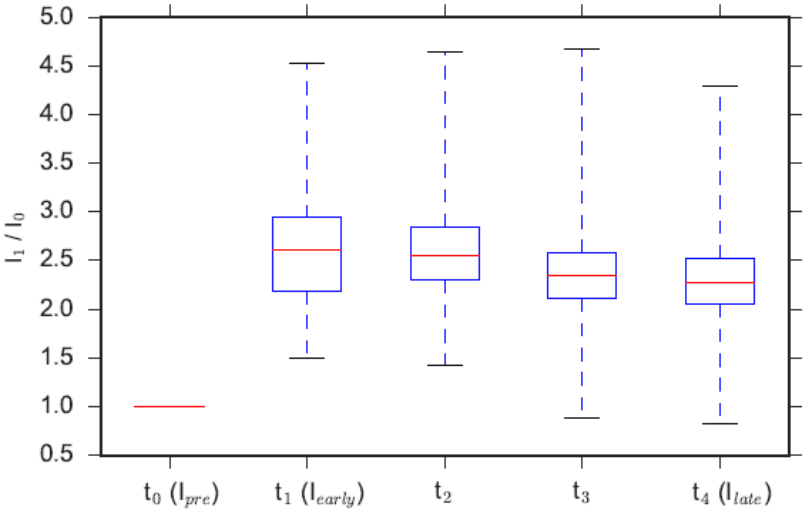
where the post-contrast images were registered to the pre-contrast image using a fully automated deformable registration (Dmitriev et al., 2013). Here, k denotes the index of a voxel in the image, I_{pre} is the voxel intensity in the pre-contrast image, I_{early} is the voxel intensity in the first post-contrast image, and I_{late} is the voxel intensity in the last post-contrast image.



(a)



(b)



(c)

Figure 1. Example of an invasive ductal carcinoma of 23 mm in a 47-year-old female patient: (a) and (b) show a coronal T1-weighted slice through the entire breast area for the washin and washout images respectively, (c) shows boxplots of the intensity ratio I_i/I_0 for the different time points of the image series, where I_i and I_0 are the voxel intensities for the i^{th} and the pre-contrast time points respectively. The voxels included for this graph are those in the tumor region of interest with initial enhancement larger than 50%, and the boxplots show the mean value, the interquartile range and the 95% confidence interval.

For the current study, masks of the tumor were available – semi-automatically determined (Alderliesten et al., 2007) and confirmed by a breast MR radiologist (C.L., with more than 10 years of experience) – to indicate the location and size of the tumor. Using these masks, a box-shaped region of interest (ROI) was established around each tumor in three orthogonal directions, encompassing the entire tumor with a margin of at least one voxel. The same ROI was extracted from the washin and washout images, each rescaled to a uniform size using trilinear interpolation. The rescaled ROI size was chosen to be 16×16×16 voxels so that the average scaling factor of the tumor volumes was 1. The washin and washout intensity values of the tumor volumes were concatenated to form the tumor's feature vector (Figure 2).

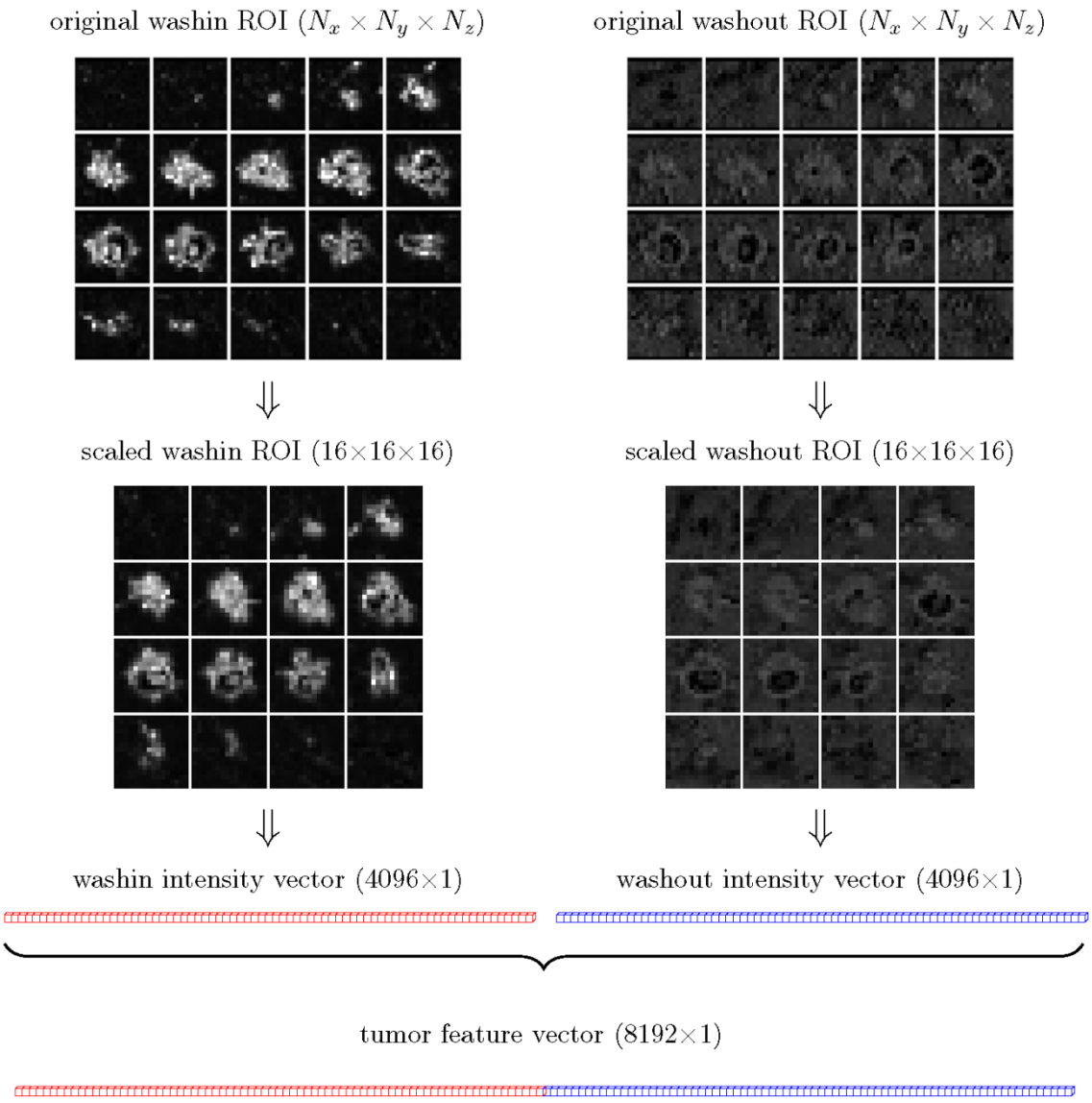


Figure 2. Schematic showing the extraction of the feature vector from a single tumor's region of interest (ROI). N_x, N_y, and N_z are the sizes of the original (non-standardized) washin and washout ROIs in the *x*-, *y*-, and *z*-direction, respectively. The ROIs are scaled to a standard volume of 16×16×16 voxels. The intensity values of the voxels in the ROIs are used to construct washin and washout intensity vectors of length 4096, and concatenating these vectors results in a tumor feature vector of length 8192.

The tumor features were used for training a model predicting treatment failure (Figure 3). The principal components – our "eigentumors" – were calculated and the projections of the feature vectors along these components were determined (i.e., the eigenvalues). Only the eigentumors that explained 90% of the total data variance were considered as candidate predictors for overall survival ("yes" or "no") at a median of 86 months by the LASSO. The regularization parameter of LASSO was set to

reduce the number of components, with a maximal reduction of 95%. The remaining components were subsequently fitted to predict overall survival using logistic regression.

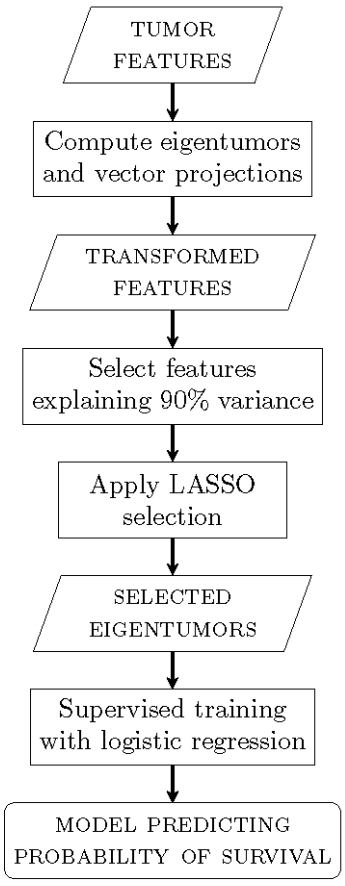


Figure 3. The steps for constructing the prediction model from the tumor feature vectors. From the feature vectors, the eigentumors are computed. The original features are projected along these eigentumors to obtain transformed features. A least absolute shrinkage selection operator (LASSO) is used to select relevant eigentumors which explain 90% of the data's variance. The selected eigentumors are used to train a model which outputs the probability of survival.

Internal validation of the predictive model

Receiver operating characteristic (ROC) curves were used to assess the performance of the model. The area under the curve (AUC) was computed for the training analysis (i.e., training and testing on the same data set), as well as for bootstrapping and leave-one-out cross-validation (LOOCV). Bootstrapping was performed with 1000 bootstrap cycles with bootstrap sample size equal to the number of cases in the dataset. To gauge the probability of overfitting with the chosen regularization

parameter for the LASSO, the regularization parameter was varied over a wider range of values and the corresponding AUCs were computed.

The performance of the model was further confirmed with Kaplan-Meier survival curves, which were computed as follows. Each bootstrap cycle yields predicted event probabilities for the test cases of that cycle. The test cases over all bootstrap cycles were randomly partitioned into subsets, so that none of the sets contained more than one prediction instance per case. The cases in each subset were then divided into a high risk group and a low risk groups based on their predicted probability value, using a threshold of 0.5. With the survival events and censoring times for the dataset, Kaplan-Meier survival curves were determined and a log-rank test for significance was applied to each subset. Finally, the median values of the resulting χ-values and corresponding *P*-values were computed over all subsets. The median survival curve and 95% confidence interval were also plotted per prediction group. This survival analysis was performed for all cases and for subgroups based on tumor subtype, and, additionally, OS hazard ratios were computed. R version 3.3.0 (R Foundation for Statistical Computing, Vienna, Austria) was used for the Kaplan-Meier survival statistics and Cox regression.

We investigated whether treatment differed significantly between patients considered to be at low risk versus those considered to be at high risk according to the prediction model. For this purpose, the patient cases were stratified in four groups (no treatment, chemotherapy, hormone therapy or both). Chi-squared testing between low- and high-risk groups was used to assess differences in assigned therapy.

The effect of noise on the model performance was also investigated. We simulated noisy images by adding Gaussian noise to the washin and washout images. The added noise had a mean value of 0, a variance of σ^2, and several values of σ were chosen. The predictive model trained on original washin and washout images was then evaluated on the images with noise added. The area-under-the-curve values were computed and compared for the different noise levels.

Results

Patient cohort

A total of 563 patients were included. The patient age at diagnosis ranged from 26 to 84 years, with median of 57.3 years. Age at diagnosis, follow-up time and number of overall survival events showed no significant difference between immunohistochemical subtypes. The variables that differed significantly between the IHC groups, with $P < 0.0001$, were largest tumor diameter, histologic tumor grade, circularity, irregularity, washin, and treatment received (Table 1). For ten cases, the IHC subtype could not be determined due to missing expressions for the hormone receptors.

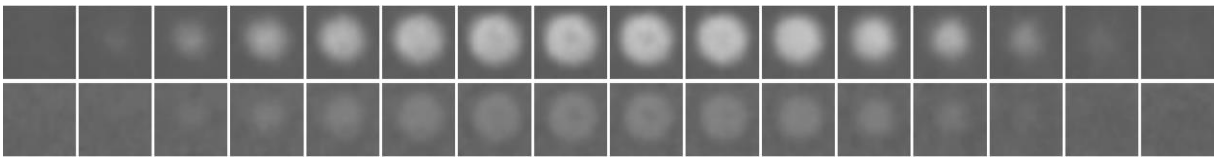
Table 1. Characteristics of the patient cohort. Values between parentheses are percentages, values between square brackets indicate ranges.

	All (*n*=563)	ER-positive, HER2-negative (*n*=408)	HER2-positive, ER-negative (*n*=73)	Triple-negative (*n*=72)	Not known (*n*=10)	*P*-value
Median age at diagnosis (years)	57.3 [26 – 84]	57.5 [26 – 84]	56.8 [29 – 69]	55.2 [27 – 77]	54.0 [40 – 63]	0.0946
Median largest tumor diameter (mm)	19 [5 – 90]	17 [5 – 90]	20 [8 – 73]	23 [5 – 60]	36 [9 – 39]	<0.0001
Median circularity	0.78 [0.31 – 0.99]	0.78 [0.31 – 0.99]	0.76 [0.45 – 0.91]	0.79 [0.33 – 0.90]	0.60 [0.44 – 0.86]	<0.0001
Median irregularity	0.42 [0.25 – 0.75]	0.41 [0.25 – 0.75]	0.45 [0.32 – 0.75]	0.47 [0.31 – 0.74]	0.59 [0.35 – 0.70]	<0.0001
Tumor voxel fraction	0.22 [0.07 – 0.35]	0.22 [0.07 – 0.35]	0.21 [0.10 – 0.31]	0.22 [0.14 – 0.33]	0.18 [0.13 – 0.21]	0.0042
Median uptake speed	1.62 [0.00 – 4.41]	1.57 [0.00 – 3.28]	1.71 [0.79 – 4.41]	1.72 [0.42 – 3.94]	1.22 [0.58 – 2.00]	0.0022
Median washout	0.15 [-0.27 – 0.42]	0.15 [-0.27 – 0.37]	0.17 [-0.17 – 0.42]	0.18 [-0.27 – 0.39]	0.05 [-0.09 – 0.24]	0.0137
Median time to follow up (months)	86 [3 – 150]	85 [5 – 148]	57 [38 – 148]	86 [3 – 150]	78.5 [53 – 119]	0.4725
Overall survival						0.1023
Event occurred	53 (9.4)	36 (8.8)	5 (6.8)	12 (16.7)	0 (0.0)	
Censored	510 (90.6)	372 (91.2)	68 (93.2)	60 (83.3)	10 (100.0)	
Histologic grade						<0.0001
Grade 1	177 (31.4)	168 (41.2)	3 (4.1)	5 (6.9)	1 (10.0)	
Grade 2	233 (41.4)	185 (45.3)	31 (42.5)	10 (13.9)	7 (70.0)	
Grade 3	142 (25.2)	47 (11.5)	38 (52.0)	55 (76.4)	2 (20.0)	
Missing	11 (2.0)	8 (2.0)	1 (1.4)	2 (2.8)	0 (0.0)	
Number of positive lymph nodes						0.0071
No positive lymph nodes	370 (65.7)	275 (67.6)	36 (49.3)	51 (70.9)	8 (80.0)	
1-3 positive lymph nodes	152 (27.0)	108 (26.5)	29 (39.7)	15 (20.8)	0 (0.0)	
Four or more positive lymph nodes	35 (6.2)	23 (5.7)	6 (8.2)	6 (8.3)	0 (0.0)	
Missing	5 (1.1)	1 (0.2)	2 (2.8)	0 (0.0)	2 (20.0)	
Chemotherapy						<0.001

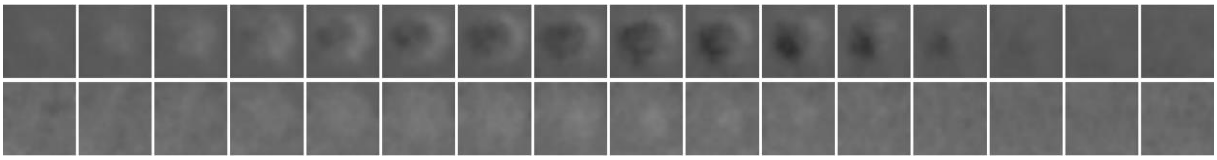
Yes	167 (29.6)	88 (21.6)	38 (52.0)	41 (56.9)	0 (0.0)	
No	395 (70.2)	320 (78.4)	34 (46.6)	31 (43.1)	10 (100.0)	
Missing	1 (0.02)	0 (0.0)	1 (1.4)	0 (0.0)	0 (0.0)	
Radiotherapy						<0.001
Yes	478 (84.9)	357 (87.5)	55 (75.3)	61 (84.7)	5 (50.0)	
No	84 (14.9)	51 (12.5)	17 (23.3)	11 (15.3)	5 (50.0)	
Missing	1 (0.2)	0 (0.0)	1 (1.4)	0 (0.0)	0 (0.0)	
Hormone therapy						<0.001
Yes	215 (38.2)	179 (43.9)	35 (47.9)	1 (1.4)	0 (0.0)	
No	347 (61.6)	229 (56.1)	37 (50.7)	71 (98.6)	10 (100.0)	
Missing	1 (0.2)	0 (0.0)	1 (1.4)	0 (0.0)	0 (0.0)	

Construction of the predictive model

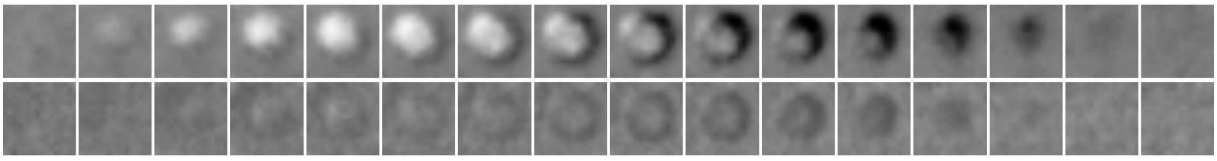
Ninety percent of the variance of the data was explained by the first 322 principal components. The LASSO with regularization parameter 3.5 selected 28 of these components. The first eigentumor shows a fairly uniform enhancement pattern, possibly resembling an average of all tumors (Figure 4(a)). Other eigentumors capture different enhancement patterns, for instance gradients in various directions (Figure 4(c), Figure 4(e) and Figure 4(f)), or differences in enhancement at the center compared to the periphery (Figure 4(d)).



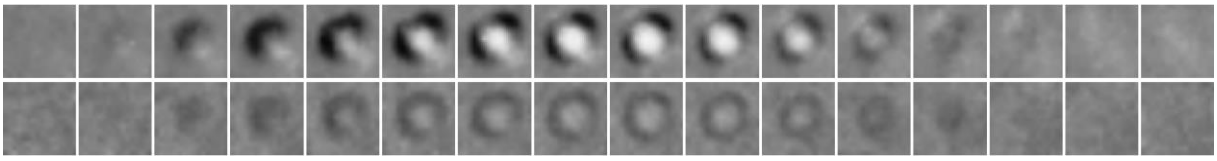
a)



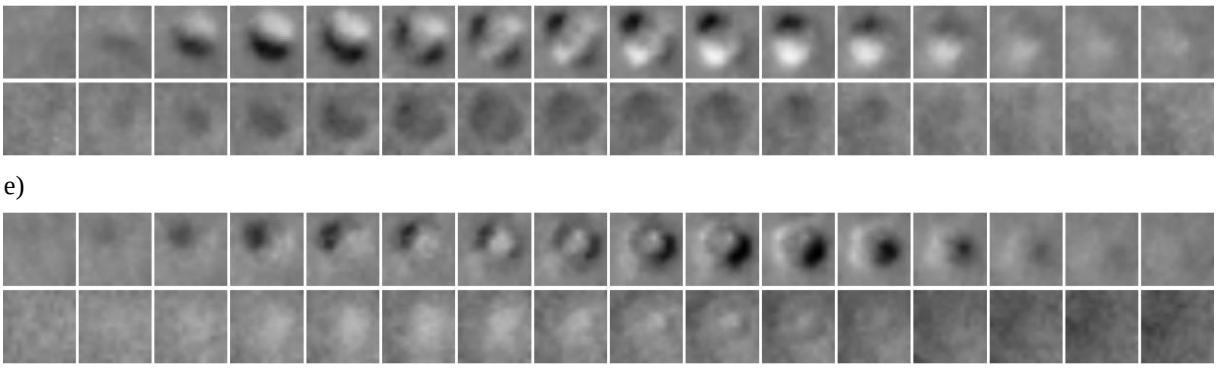
b)



c)



d)



e)

f)

Figure 4. Two-dimensional representations of the (a) first, (b) second*, (c) sixth*, (d) seventh*, (e) nineth* and (f) twelfth* eigentumors. The eigentumors are shown as slices in the axial plane for the washin intensities (top rows) and the washout intensities (bottom rows). Components marked with an asterisk (*) were among the selected candidate predictors.

Internal validation of the predictive model

The training, leave-one-out and bootstrapped AUCs were 0.88, 0.77 and 0.73 respectively (Figure 5). For the range of regularization parameter values, the leave-one-out performance values were consistently higher than the bootstrapped values (Figure 6).

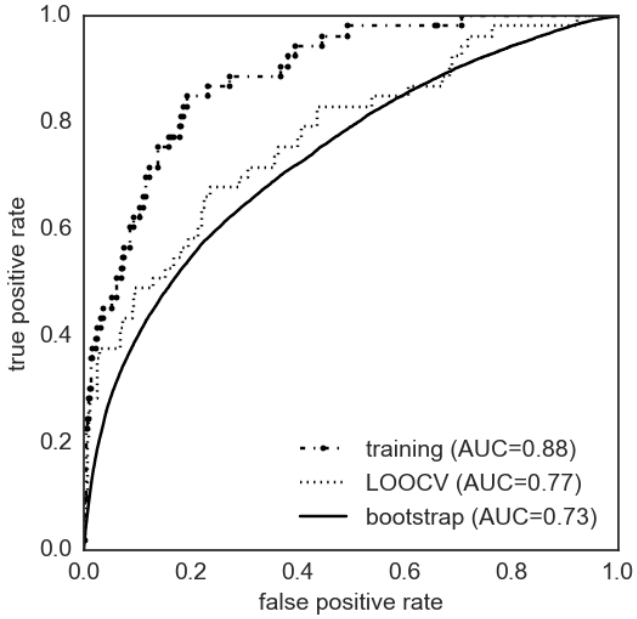


Figure 5. Receiver operating characteristic curve for overall survival at a median of 86 months. The dash-dotted curve denotes the training performance, and the dotted and solid curves denote the leave-one-out cross-validated (LOOCV) and bootstrapped performances respectively.

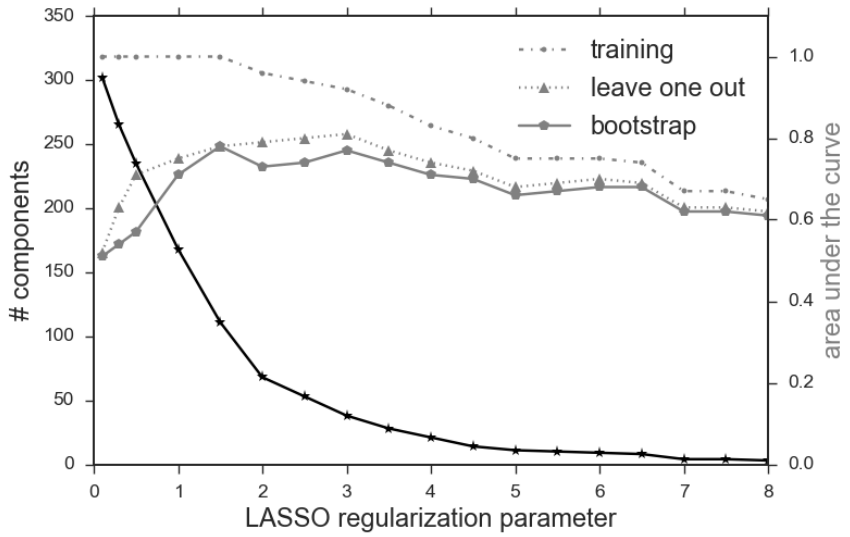
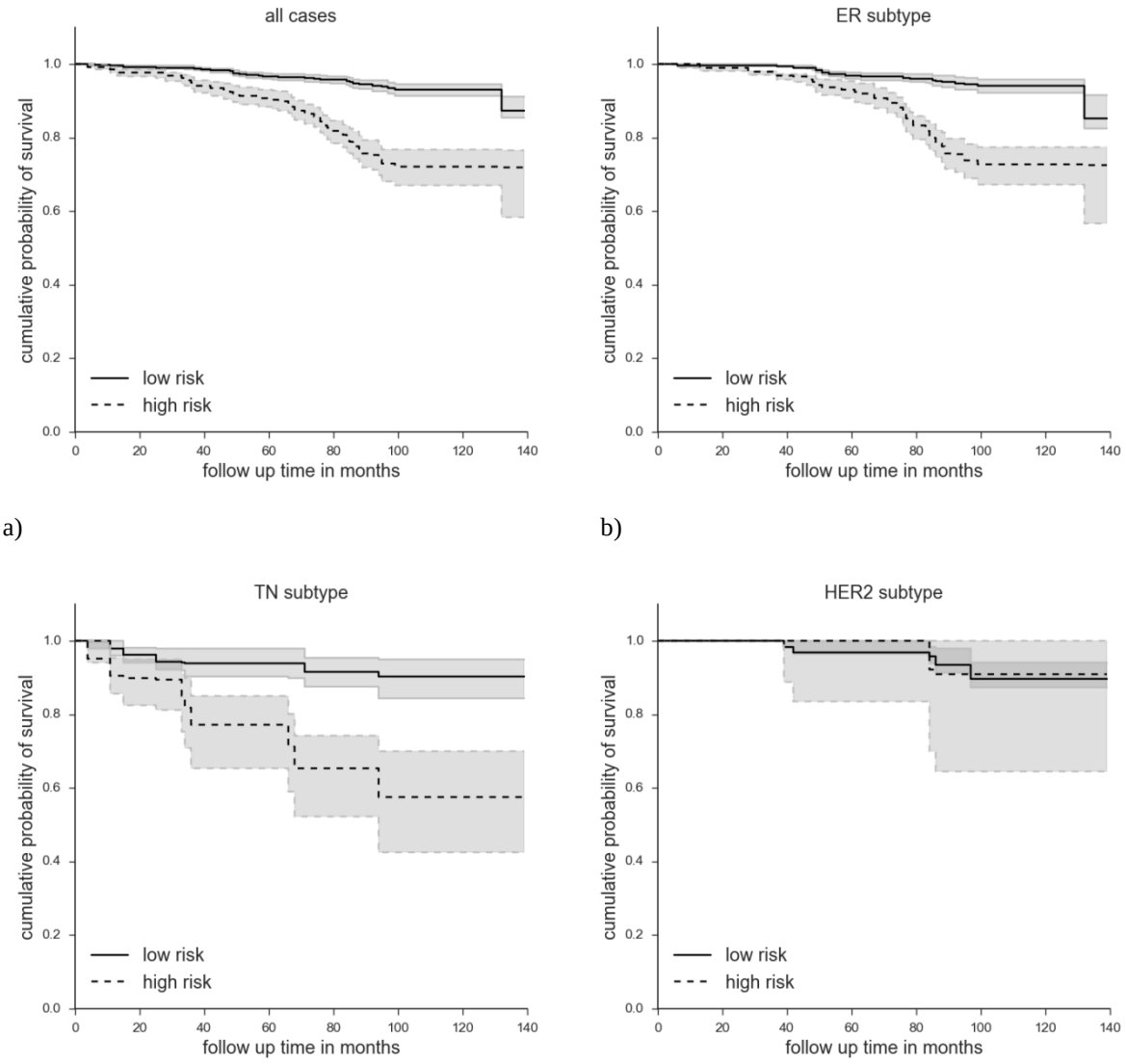


Figure 6. The number of selected components (black curve, values on left vertical axis) and areas under the receiver operating characteristic curve as measure of performance (gray curves, values on right vertical axis) plotted against the LASSO regularization parameter (horizontal axis).

Log-rank test analysis showed significant separation between the low risk and high risk patient groups when all subtypes were included: the entire range of *P*-values over the bootstrap cycles was well below the threshold for significance of 0.05, with median $P < 0.0001$ (Figure 7(a)). The high-risk group was associated with worse overall survival: the OS hazard ratio (HR) was 4.31 [2.50–7.42]. The survival analysis with subtype taken into account showed that the ER subtype, which accounted for 70% of all cases, has similar results (OS HR=4.46 [2.30 – 8.70]) (Figure 7(b)). The median *P*-values of the log-rank test were $P=0.0039$ for the TN-subtype, and $P=0.41$ for the HER2-subtype (Figure 7(c), Figure 7(d)). For comparison, when stratifying all cases by immunohistochemical subtype, no significant separation ($P=0.077$) between the survival curves is found (Figure 8).



a) b)

c) d)

Figure 7. Bootstrapped Kaplan-Meier curve stratified by probability of overall survival event (low risk versus high risk according to the eigentumor model), where the grey areas indicate the 95% confidence intervals: (a) for tumors of all subtypes (median: χ=32.89, P < 0.0001), (b) for estrogen-receptor subtype (ER-positive, HER2-negative) tumors (median: χ=23.78, P < 0.0001), (c) for triple-negative tumors (median: χ=8.30, P < 0.0039) and (d) for HER2-subtype (HER2-positive) tumors (median: χ=0.69, P=0.41).

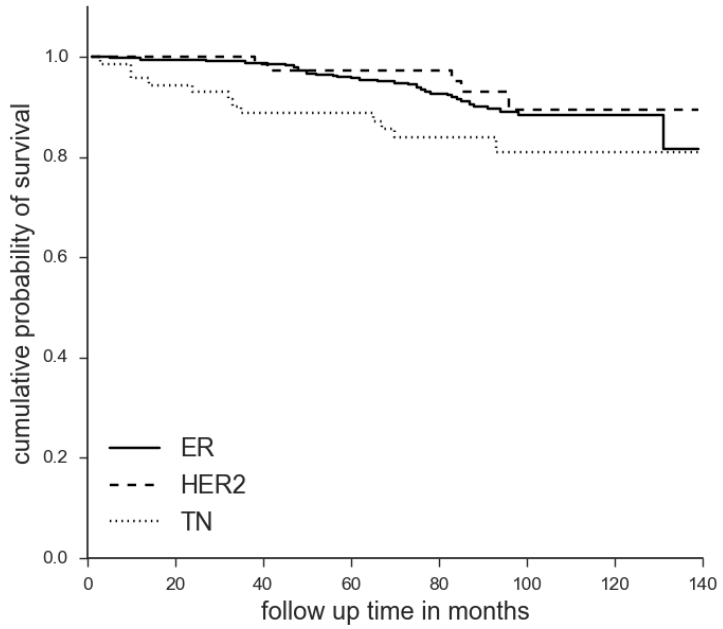


Figure 8. Kaplan-Meier survival curve for all tumors grouped by immunohistochemical subtype: ER (ER-positive, HER2-negative), HER2 (HER2-positive) and TN (triple-negative) subtype (χ=6.836, *P*=0.077).

We found no significant difference in received therapy (chemo, hormone, both, or none) of patients considered to be at low risk according to the model and those considered to be at high-risk (*P*=0.68). This suggests that the model has not been influenced by the therapy given, and that it may have complimentary value to existing prognostic and predictive markers for treatment selection.

As expected, the addition of Gaussian noise with increasing values of σ to the rescaled washin and washout images resulted in ROI images in which the tumor becomes increasingly obscured (Figure 9). Performance evaluation on the noisy images shows that the area-under-the-curve values decrease as the amount of noise is increased, but remain significantly above chance performance, with a lowest AUC-value of 0.77 for the range under investigation (Figure 10).

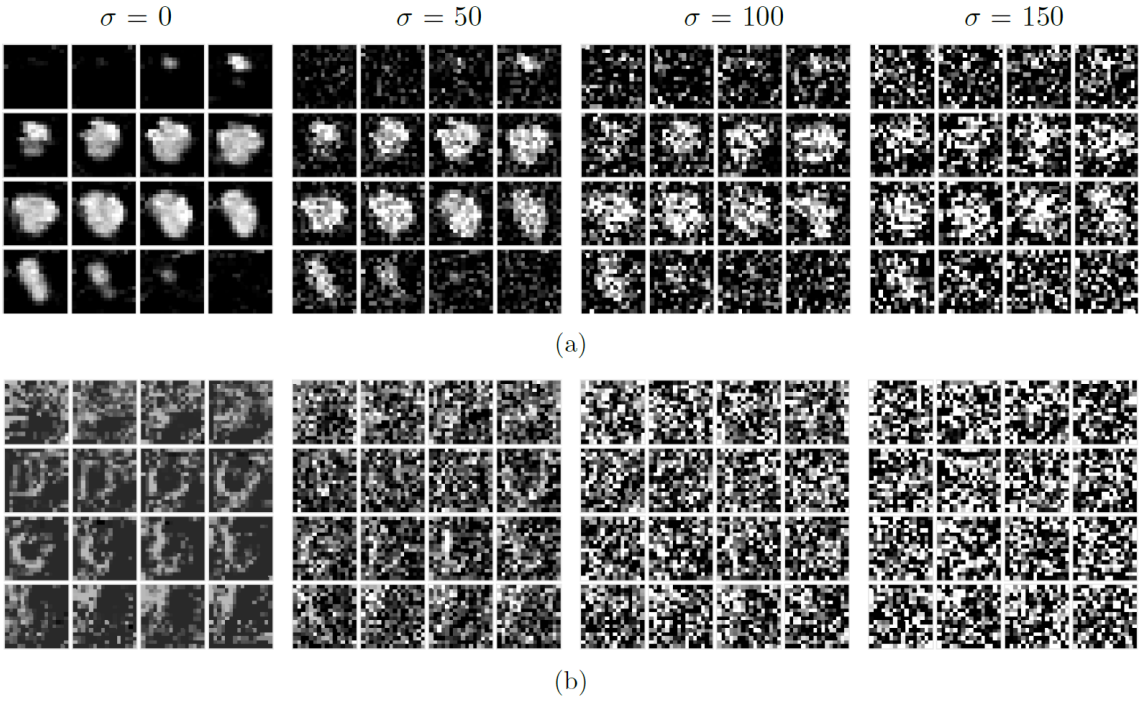


Figure 9. Examples of rescaled washin (a) and washout (b) images with varying levels of noise added. The first ROI image shows the original washin or washout image (σ=0), and the three images on the right show added noise images for σ=50, σ=100 and σ=150 respectively.

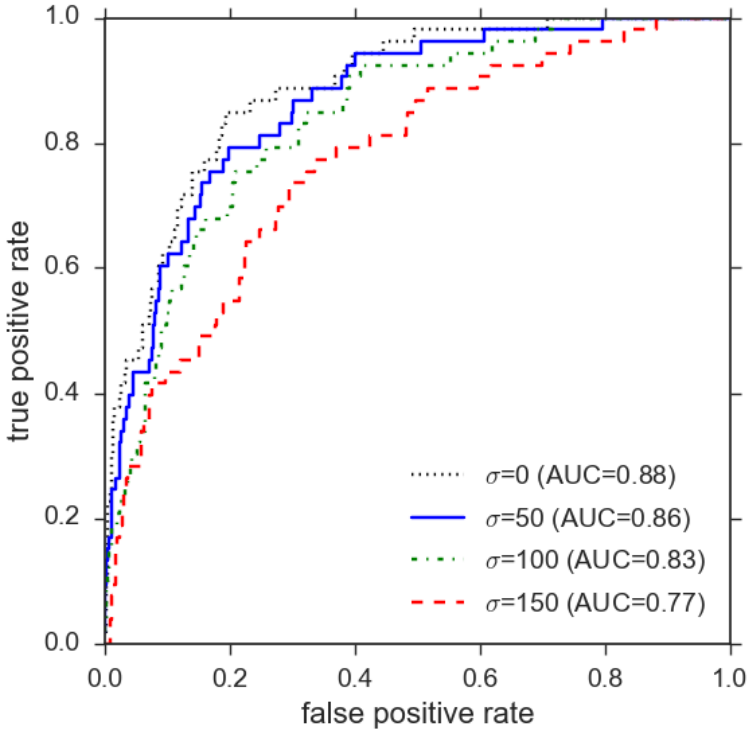


Figure 10. Receiver operating characteristic curves for model evaluation on the simulated images with increasing noise levels.

Discussion

In this study we developed a fully automated method – eigentumor analysis – to predict treatment failure in patients with early breast cancer from a single pre-treatment MRI scan. The method does not rely on pre-defined engineered features and does not require lesion segmentation. In a dataset of 563 consecutively included patients, the eigentumors were found to stratify patient survival after 140 months significantly with a hazard ratio of 4.31 [2.50–7.42].

The selection of appropriate breast cancer treatment is particularly important for patients diagnosed with cancer at an early stage. It would therefore be useful to be able to predict the probability of survival or treatment failure. With increasing number of pattern recognition tools and increasing size of medical image datasets, the extraction of quantitative image features and the analysis of these data for decision support – a research field called radiomics – has become a more frequent research practice (Gillies et al., 2016). For instance, Li et al. report that computer-extracted image phenotypes (CEIPs) show promise for assessing the risk in breast cancer recurrence (Li et al., 2016).

Image analysis is based on the hypothesis that image phenotypes may be related to a certain outcome of interest. Quantitative feature methods might enable finding relevant tumor phenotypes which are not easily distinguished by eye. Our approach uses principal component analysis (PCA) to compute eigenvectors (or eigentumors), an approach that has a number of advantages. Not only does PCA allow reduction of the dimensionality of the features, but, more importantly, the computed eigentumors are by definition independent components that contain the variance and thus the information of the full image data. This approach is in contrast to more conventional methods where one or multiple a-priori defined image features are extracted. These “engineered” features are hypothesized to be associated with the outcome, and the subset with the strongest association is selected. Although this approach works well for problems in which radiologists are able to provide input on hypothesized associations between phenotype and outcome (such as benign versus malignant

disease), it is more difficult for problems where radiologists have not yet formed such associations because they lose track of patients after their treatment is completed.

Relatively few studies with engineered features focus on survival prediction for early-stage breast cancer patients. In one such study, Kim et al. found that T1 and T2 entropy of the lesion is associated with recurrence free survival (RFS) in 203 early-stage breast cancer patients with mass lesions (Kim et al., 2017). The reported hazard ratios were of the same range as those in the current study (HR=4.31 versus 4.55 and 9.87). It is promising that both studies find evidence of image phenotypes being associated with treatment failure, despite differences in methodology used (analysis of entropy versus eigentumors, manual lesion segmentation versus no lesion segmentation) and differences in datasets (563 patients with early breast cancer without additional exclusion at the gate in the current study).

A larger number of studies investigate treatment response in the neoadjuvant setting for patients with locally advanced breast cancer, a different patient population to that of the current study. Quantitative texture features, which encompass a broad array of features derived from histograms, gray-scale correlation matrices or local binary patterns have been investigated to this end (Golden et al., 2013, Teruel et al., 2014). Other engineered imaging features reported to correlate with treatment response include pharmacokinetic parameters, such as the volume transfer coefficient (K^{trans}) (Ah-See et al., 2008, Golden et al., 2013, He et al., 2012), as well as diffusion related features such as the apparent diffusion coefficient (ADC) (Park et al., 2010, Richard et al., 2013).

Other combinations of principal component analysis with breast DCE-MRI have previously been reported for applications such as tumor segmentation (Agner et al., 2013, Akhbardeh and Jacobs, 2012), classification of tumor benignity or malignity (Eyal et al., 2009, Furman-Haran et al., 2014, Levman et al., 2014), and prediction of pathological response after neoadjuvant chemotherapy (Wu et al., 2016). However, the principal components of the methods previously reported were computed over the kinetic curve of a single voxel (or an average curve over multiple voxels) inside a tumor. Thus, feature vectors contain intensity values at consecutive times, and principal components describe the temporal contrast uptake of the tumors only. In contrast, the principal components constructed as in the current study contain both intensity and (three-dimensional) morphological information of the

tumor, because the feature vectors combine the voxel intensities of the washin and washout in and around the tumor while maintaining the spatial interrelationship between the voxels.

For principal component analysis, input feature vectors are required to have equal lengths. To arrive at uniformly sized rescaled ROIs, the regions of interest around the tumor were chosen to be box-shaped, where such a box also contains non-tumor tissue surrounding the tumor. Studies have shown stromal volume surrounding lesions to be associated with response to neoadjuvant therapy (Hattangadi et al., 2008), thus we expect that taking the surrounding non-tumor tissue into account will be useful as well. Since tumor masks were available for this dataset, these were used to determine the region of interest around the tumor. However, the principal component extraction method itself does not depend on the shape of the tumor segmentations. The input of the model is a box-shaped region of interest containing the tumor, which can also be obtained without complete segmentation, be it automatically, for a fully automated workflow, or manually by a radiologist. We do not expect that our results will be influenced greatly by the method of ROI selection, as long as the ratios between the number of tumor voxels and non-tumor voxels in the region of interest are more or less consistent with the cases on which the model is trained.

We chose to keep the principal components that explained 90% of the variance of the data, a commonly used threshold. Other options would be to retain the components with eigenvalues larger than a certain threshold eigenvalue (e.g. the average eigenvalue), or the eigenvalue at the "elbow point" in the PCA's scree plot. We used the method that discarded the fewest components, because the least absolute shrinkage selection operator (LASSO) is subsequently used, performing a robust variable selection by regularization.

The LASSO is a method for finding good predictors in data of high dimensionality (Greenshtein, 2006). However, the regularization parameter for the LASSO can be very influential on the outcome. When choosing a regularization parameter, sufficiently large datasets are usually split into a training set and a test set. The LASSO's regularization parameter is then determined by performing cross-validation on the training set (usually ten-fold or leave-one-out cross-validation), after which validation of the trained model is performed on the unseen data of the test set. For smaller studies, the regularization parameter is often chosen based on the cross-validated performance, without

separate test-set validation. However, this is likely to yield estimated prospective performance values that are too optimistic, particularly in the case of leave-one-out cross-validation (LOOCV). Splitting the dataset into a training set and a test set also has some drawbacks. In particular, the model is trained on a smaller portion of the data, which can affect the learning performance adversely. Furthermore, the resulting test performance might be influenced by the specific splits into training and test set. Bootstrapping allows averaging over many instances of selected training and test sets, and is thus expected to be less sensitive to the specific set divisions. A simulation study by Breiman and Spector also shows that bootstrapping performs at least as good as, or better than, ten-fold cross-validation for submodel selection and evaluation (Breiman and Spector, 1992). Therefore, we performed internal cross-validation using bootstrapping. We also determined performance values for leave-one-out cross-validation, because other studies often report LOOCV performance values, and this may enable comparisons.

In the current study, we chose an empirical regularization value for which we assessed both training and internally cross-validated performances. The effect of varying the regularization parameter on the performance was investigated subsequently. Without regularization, all components are used for fitting. This yields cross-validated performances which are only slightly above 0.5, while the training performance is close to 1.0. Such a discrepancy between training and cross-validation performances indicates that the model has overfitted to the training data and does not generalize well to unseen data. Applying regularization – and thus, selecting fewer components for fitting – reduces the difference between the performances: the cross-validated performances are improved, but the training performance also decreases. The bootstrapped and leave-one-out cross-validated performances show a clear increase when the regularization parameter is increased from 0 to 1. In the parameter range between 1 and 6, these cross-validated performances remain fairly stable and do not drop below 0.6, even if the regularization is increased further. Therefore, we believe that the regularization parameter we chose (3.5) is a reasonable value, with which we expect to have avoided overfitting. The results from investigating the addition of noise to the image dataset seem to confirm this expectation.

Stratifying the tumors of our dataset on immunohistochemical subtype shows a trend between tumor subtype and survival (Figure 8). However, the effect of subtype on patient survival might be obscured, because the patients' treatments were decided partly based on this IHC subtype. Regardless of this optimal chosen treatment based on all available information at the time, the presented eigentumor model seems to be able to discriminate between low and high risk groups for treatment failure for tumors of ER and TN-subtype.

The bootstrapped stratification into high- and low-risk groups splits the patient survival significantly, but in a number of cases the survival is not predicted correctly (AUC=0.73). To verify that our model predictions do not suffer from systematic errors, we studied whether patient- and tumor characteristics (i.e., largest tumor diameter, time to follow up, number of positive lymph nodes, age at diagnosis), differ significantly between cases in which survival was predicted correctly and those in which it was not. We did not find such differences. This seems to indicate that errors in model predictions are mostly stochastic in nature and related to the data sample size, the degrees of freedom in the model, and the classifier. Therefore, adding more (and more diverse) cases to our training dataset, combined with other, more advanced, classifiers seems to be the most promising approach to improve the model presented here.

Our study also has limitations. All MR image data were obtained from a single institution, using the same imaging unit. Differences in MR field strength, manufacturer, imaging protocols, and contrast agents may have influence on our model. These factors may be investigated in a multi-institutional retrospective cohort study in which sufficiently long follow-up data is available. Imaging protocols typically range from "fast" series which have emphasis on temporal resolution to "slow" series with emphasis on spatial resolution. We anticipate that balance between these extremes will be most effective for the accuracy of the model.

Robust and practical diffusion-weighted imaging (DWI) protocols and fast dynamic protocols were not yet in place at the Netherlands Cancer Institute at the time of patient inclusion. Therefore, the MARGINS dataset did not contain diffusion-weighted images or fast dynamic scans, and neither apparent diffusion coefficient maps nor pharmacokinetic features could be computed (Henderson et al., 1998). These limitations came with the advantage of having access to long-term follow-up information.

Despite the absence of these imaging protocols, the eigentumors based on conventional and still widely used DCE-MRI protocols already stratified patient survival significantly.

If necessary, the imaging protocol may be standardized across institutes for a particular prognostic test. The currently used protocol is based on such balance and is still widely applied in many breast MR examinations. Another subject for follow-up study is to investigate the potential complementary value of the test with respect to currently available prognostic models. Future research could also include incorporating pathological prognostic factors that are known to be related to survival into the model. Because the model stratified survival into a high-risk and a low-risk group while treatment was not assigned differently in these groups, we anticipate that the model – once validated – will show complementary value to routine markers that are currently used for treatment selection.

Conclusions

The performance of the presented model applied to DCE MRI indicates that selected eigentumors – principal components based on the temporal and morphological characteristics of tumors – have potential for predicting treatment failure in early breast cancer patients. We are aware that this is primarily a proof-of-concept study, and that the eigentumor model may be further validated on an independent dataset.

Acknowledgements

This work is part of the research program IMDI with project number 104003019, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

Disclosure of Conflicts of Interest

The authors have no relevant conflicts of interest to disclose.

References

AGNER, S. C., XU, J. & MADABHUSHI, A. 2013. Spectral embedding based active contour (SEAC) for lesion segmentation on breast dynamic contrast enhanced magnetic resonance imaging. *Medical Physics,* 40**,** 032305.

AH-SEE, M. L., MAKRIS, A., TAYLOR, N. J., HARRISON, M., RICHMAN, P. I., BURCOMBE, R. J., STIRLING, J. J., D'ARCY, J. A., COLLINS, D. J., PITTAM, M. R., RAVICHANDRAN, D. & PADHANI, A. R. 2008. Early changes in functional dynamic magnetic resonance imaging predict for pathologic response to neoadjuvant chemotherapy in primary breast cancer. *Clin Cancer Res,* 14**,** 6580-6589.

AKHBARDEH, A. & JACOBS, M. A. 2012. Comparative analysis of nonlinear dimensionality reduction techniques for breast MRI segmentation. *Medical Physics,* 39**,** 2275-2289.

ALDERLIESTEN, T., SCHLIEF, A., PETERSE, J., LOO, C., TEERTSTRA, H., MULLER, S. & GILHUIJS, K. 2007. Validation of semiautomatic measurement of the extent of breast tumors using contrast-enhanced magnetic resonance imaging. *Investigative Radiology,* 42**,** 42-49.

BREIMAN, L. & SPECTOR, P. 1992. Submodel Selection and Evaluation in Regression - the X-Random Case. *International Statistical Review,* 60**,** 291-319.

CHEN, X. S., MA, C. D., WU, J. Y., YANG, W. T., LU, H. F., WU, J., LU, J. S., SHAO, Z. M., SHEN, Z. Z. & SHEN, K. W. 2010. Molecular subtype approximated by quantitative estrogen receptor, progesterone receptor and Her2 can predict the prognosis of breast cancer. *Tumori,* 96**,** 103-110.

CIANFROCCA, M. & GOLDSTEIN, L. J. 2004. Prognostic and predictive factors in early-stage breast cancer. *Oncologist,* 9**,** 606-616.

DE CAMARGO MORAES, P., CHALA, L. F., CHANG, Y. S., KIM, S. J., ENDO, E., DE BARROS, N. & SPINOLA, F. 2010. Observer variability in the application of morphologic and dynamic criteria according to the BI-RADS for MRI. *Breast J,* 16**,** 558-560.

DE MASCAREL, I., DEBLED, M., BROUSTE, V., MAURIAC, L., SIERANKOWSKI, G., VELASCO, V., CROCE, S., CHIBON, F., BOUDEAU, J., DEBANT, A. & MACGROGAN, G. 2015. Comprehensive prognostic

analysis in breast cancer integrating clinical, tumoral, micro-environmental and immunohistochemical criteria. *Springerplus,* 4**,** 528.

DMITRIEV, I. D., LOO, C. E., VOGEL, W. V., PENGEL, K. E. & GILHUIJS, K. G. 2013. Fully automated deformable registration of breast DCE-MRI and PET/CT. *Phys Med Biol,* 58**,** 1221-1233.

EYAL, E., BADIKHI, D., FURMAN-HARAN, E., KELCZ, F., KIRSHENBAUM, K. J. & DEGANI, H. 2009. Principal Component Analysis of Breast DCE-MRI Adjusted With a Model-Based Method. *Journal of Magnetic Resonance Imaging,* 30**,** 989-998.

FOCKE, C. M., DECKER, T. & VAN DIEST, P. J. 2016. Reliability of histological grade in breast cancer core needle biopsies depends on biopsy size: a comparative study with subsequent surgical excisions. *Histopathology,* 10.1111/his.13036**,** 10.1111/his.13036.

FURMAN-HARAN, E., FEINBERG, M. S., BADIKHI, D., EYAL, E., ZEHAVI, T. & DEGANI, H. 2014. Standardization of Radiological Evaluation of Dynamic Contrast Enhanced MRI: Application in Breast Cancer Diagnosis. *Technology in Cancer Research & Treatment,* 13**,** 445-454.

GILHUIJS, K. G. A., DEURLOO, E. E., MULLER, S. H., PETERSE, J. L. & SCHULTZE KOOL, L. J. 2002. Breast MR imaging in women at increased lifetime risk of breast cancer: clinical system for computerized assessment of breast lesions—initial results. *Radiology,* 225**,** 907-916.

GILHUIJS, K. G. A., GIGER, M. L. & BICK, U. 1998. Computerized analysis of breast lesions in three dimensions using dynamic magnetic resonance imaging. *Medical Physics,* 25**,** 1647-1654.

GILLIES, R. J., KINAHAN, P. E. & HRICAK, H. 2016. Radiomics: Images Are More than Pictures, They Are Data. *Radiology,* 278**,** 563-577.

GOLDEN, D. I., LIPSON, J. A., TELLI, M. L., FORD, J. M. & RUBIN, D. L. 2013. Dynamic contrast-enhanced MRI-based biomarkers of therapeutic response in triple-negative breast cancer. *Journal of the American Medical Informatics Association,* 20**,** 1059-1066.

GREENSHTEIN, E. 2006. Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *The Annals of Statistics,* 34**,** 2367-2386.

HAMY-PETIT, A. S., BELIN, L., BONSANG-KITZIS, H., PAQUET, C., PIERGA, J. Y., LEREBOURS, F., COTTU, P., ROUZIER, R., SAVIGNONI, A., LAE, M. & REYAL, F. 2016. Pathological complete response and prognosis after neoadjuvant chemotherapy for HER2-positive breast cancers before and after trastuzumab era: results from a real-life cohort. *Br J Cancer,* 114**,** 44-52.

HATTANGADI, J., PARK, C., REMBERT, J., KLIFA, C., HWANG, J., GIBBS, J. & HYLTON, N. 2008. Breast stromal enhancement on MRI is associated with response to neoadjuvant chemotherapy. *AJR Am J Roentgenol,* 190**,** 1630-1636.

HE, D. F., MA, D. Q. & JIN, E. H. 2012. Dynamic Breast Magnetic Resonance Imaging: Pretreatment Prediction of Tumor Response to Neoadjuvant Chemotherapy. *Clinical Breast Cancer,* 12**,** 94-101.

HENDERSON, E., RUTT, B. K. & LEE, T. Y. 1998. Temporal sampling requirements for the tracer kinetics modeling of breast disease. *Magn Reson Imaging,* 16**,** 1057-1073.

HUDIS, C. A., BARLOW, W. E., COSTANTINO, J. P., GRAY, R. J., PRITCHARD, K. I., CHAPMAN, J. A. W., SPARANO, J. A., HUNSBERGER, S., ENOS, R. A., GELBER, R. D. & ZUJEWSKI, J. A. 2007. Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: The STEEP system. *Journal of Clinical Oncology,* 25**,** 2127-2132.

JONES, E., OLIPHANT, E. & PETERSON, P. 2001. SciPy: Open source scientific tools for Python.

KIM, J. H., KO, E. S., LIM, Y., LEE, K. S., HAN, B. K., KO, E. Y., HAHN, S. Y. & NAM, S. J. 2017. Breast Cancer Heterogeneity: MR Imaging Texture Analysis and Survival Outcomes. *Radiology,* 282**,** 665-675.

LEVMAN, J., WARNER, E., CAUSER, P. & MARTEL, A. 2014. Semi-Automatic Region-of-Interest Segmentation Based Computer-Aided Diagnosis of Mass Lesions from Dynamic Contrast-Enhanced Magnetic Resonance Imaging Based Breast Cancer Screening. *Journal of Digital Imaging,* 27**,** 670-678.

LI, H., ZHU, Y., BURNSIDE, E. S., DRUKKER, K., HOADLEY, K. A., FAN, C., CONZEN, S. D., WHITMAN, G. J., SUTTON, E. J., NET, J. M., GANOTT, M., HUANG, E., MORRIS, E. A., PEROU, C. M., JI, Y. & GIGER, M. L. 2016. MR imaging radiomics signatures for predicting the risk of breast cancer

recurrence as given by research versions of MammaPrint, Oncotype DX, and PAM50 gene assays. *Radiology,* 281**,** 382-391.

MANN, R. M., KUHL, C. K., KINKEL, K. & BOETES, C. 2008. Breast MRI: guidelines from the European Society of Breast Imaging. *Eur Radiol,* 18**,** 1307-1318.

PARK, S. H., MOON, W. K., CHO, N., SONG, I. C., CHANG, J. M., PARK, I. A., HAN, W. & NOH, D. Y. 2010. Diffusion-weighted MR Imaging: Pretreatment Prediction of Response to Neoadjuvant Chemotherapy in Patients with Breast Cancer. *Radiology,* 257**,** 56-63.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. & DUCHESNAY, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* 12**,** 2825-2830.

PENNISI, A., KIEBER-EMMONS, T., MAKHOUL, I. & HUTCHINS, L. 2016. Relevance of Pathological Complete Response after Neoadjuvant Therapy for Breast Cancer. *Breast Cancer (Auckl),* 10**,** 103-106.

RAKHA, E. A., EL-SAYED, M. E., LEE, A. H., ELSTON, C. W., GRAINGE, M. J., HODI, Z., BLAMEY, R. W. & ELLIS, I. O. 2008. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. *J Clin Oncol,* 26**,** 3153-3158.

RICHARD, R., THOMASSIN, I., CHAPELLIER, M., SCEMAMA, A., DE CREMOUX, P., VARNA, M., GIACCHETTI, S., ESPIE, M., DE KERVILER, E. & DE BAZELAIRE, C. 2013. Diffusion-weighted MRI in pretreatment prediction of response to neoadjuvant chemotherapy in patients with breast cancer. *Eur Radiol,* 23**,** 2420-2431.

RICHTER-EHRENSTEIN, C., MULLER, S., NOSKE, A. & SCHNEIDER, A. 2009. Diagnostic accuracy and prognostic value of core biopsy in the management of breast cancer: a series of 542 patients. *Int J Surg Pathol,* 17**,** 323-326.

SCHMITZ, A. M., OUDEJANS, J. J. & GILHUIJS, K. G. 2014. Agreement on indication for systemic therapy between biopsied tissue and surgical excision specimens in breast cancer patients. *PLoS One,* 9**,** e91439.

SØRLIE, T., PEROU, C. M., TIBSHIRANI, R., AAS, T., GEISLER, S., JOHNSEN, H., HASTIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., THORSEN, T., QUIST, H., MATESE, J. C., BROWN, P. O., BOTSTEIN, D., LONNING, P. E. & BORRESEN-DALE, A. L. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A,* 98**,** 10869-10874.

STOUTJESDIJK, M. J., FUTTERER, J. J., BOETES, C., VAN DIE, L. E., JAGER, G. & BARENTSZ, J. O. 2005. Variability in the description of morphologic and contrast enhancement characteristics of breast lesions on magnetic resonance imaging. *Invest Radiol,* 40**,** 355-362.

TERUEL, J. R., HELDAHL, M. G., GOA, P. E., PICKLES, M., LUNDGREN, S., BATHEN, T. F. & GIBBS, P. 2014. Dynamic contrast-enhanced MRI texture analysis for pretreatment prediction of clinical and pathological response to neoadjuvant chemotherapy in patients with locally advanced breast cancer. *Nmr in Biomedicine,* 27**,** 887-896.

TIBSHIRANI, R. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological,* 58**,** 267-288.

TURK, M. & PENTLAND, A. 1991. Eigenfaces for recognition. *J Cogn Neurosci,* 3**,** 71-86.

VON MINCKWITZ, G. & FONTANELLA, C. 2015. Comprehensive Review on the Surrogate Endpoints of Efficacy Proposed or Hypothesized in the Scientific Community Today. *J Natl Cancer Inst Monogr,* 2015**,** 29-31.

WEDEGARTNER, U., BICK, U., WORTLER, K., RUMMENY, E. & BONGARTZ, G. 2001. Differentiation between benign and malignant findings on MR-mammography: usefulness of morphological criteria. *Eur Radiol,* 11**,** 1645-1650.

WU, J., GONG, G., CUI, Y. & LI, R. 2016. 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**,** 1107-1115.