A novel and fully automated mammographic texture analysis for risk prediction: results from two case-control studies

Chao Wang[1]*, Adam R Brentnall[1], Jack Cuzick[1], Elaine F Harkness[2], D Gareth Evans[3] and Susan Astley[2]

* corresponding author

[1] Centre for Cancer Prevention
Wolfson Institute of Preventive Medicine
Queen Mary University of London
Charterhouse Square
London
United Kingdom
EC1M 6BQ

[2] Centre for Imaging Science
School of Health Sciences
University of Manchester
Stopford Building
Oxford Road
Manchester
M13 9PT
United Kingdom

[3] Department of Genomic Medicine
University of Manchester
St Mary's Hospital
Manchester
M13 9WL
United Kingdom

Author email addresses:

- Chao Wang: chao.wang@qmul.ac.uk
- Adam Brentnall: a.brentnall@qmul.ac.uk
- Jack Cuzick: j.cuzick@qmul.ac.uk
- Elaine Harkness: elaine.f.harkness@manchester.ac.uk
- D Gareth Evans: gareth.evans@cmft.nhs.uk
- Susan Astley: sue.astley@manchester.ac.uk

Abstract

Background: The percentage of mammographic dense tissue (PD) is an important risk factor for breast cancer, and there is some evidence that texture features may further improve predictive ability. However, relatively little work has assessed or validated textural feature algorithms using raw full field digital mammograms (FFDM).

Method: A case-control study nested within a screening cohort (age 46-73y) from Manchester UK was used to develop a texture feature risk score (264 cases diagnosed at the same time as mammogram of contralateral breast, 787 controls) using the least absolute shrinkage and selection operator (LASSO) method for 112 features, and validated in a second case-control study from the same cohort but with cases diagnosed after the index mammogram (317 cases, 931 controls). Predictive ability was assessed using deviance and matched concordance index (mC). The ability to improve risk estimation beyond percent volumetric density (Volpara) was evaluated using conditional logistic regression.

Results: The strongest features identified in the training set were "Sum Average" based on the grey-level co-occurrence matrix at low image resolutions (original resolution 10.628 pixels per mm; downsized by factors of 16, 32 and 64), which had lower deviance and higher mC than volumetric PD. In the validation study, the risk score combining the three Sum Average features achieved a lower deviance than volumetric PD ($\Delta\chi^2$=10.55 or 6.95 if logarithm PD) as well as a similar mC to volumetric PD (0.58 and 0.57 respectively). The risk score added independent information to volumetric PD ($\Delta\chi^2$ = 14.38, p = 0.0008).

Conclusion: Textural features based on digital mammograms improve risk assessment beyond volumetric percentage density. The features and risk score developed needs further investigation in other settings.

Key words: Breast density, texture, digital mammogram, risk prediction, breast cancer

1 Background

Mammographic density is a term used to describe whiter regions of the images that reflect the amount of fibroglandular as opposed to fatty tissue in the breast. Mammographic density is a well-established risk factor for breast cancer [1]. One well studied measure of breast density is the percentage of the breast area which is opaque, often referred to as percent density (PD). In addition to area-based PD, volumetric measures have been developed to make use of the greyscale pixel values, without thresholding. It has been estimated that 16% of all breast cancers and 26% of breast cancers in women aged 55 or less years are attributable to breast density over 50% [2]; women with PD over 75% have been consistently reported to be at a four to six fold higher risk of developing the disease than women of similar age with little or no dense tissue [3]; and PD has been described as a risk factor that is the most significant after age [4].

While PD is an important risk factor, it is likely that characteristics of the mammogram other than PD may be related to breast cancer. For example, Wolfe's parenchymal patterns [5] indicate *texture* characteristics which are not necessarily correlated with PD [6]. Similarly, the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) classifies density into categories based on not only the amount of density but on descriptors of the distribution, such as "scattered" and "heterogeneously dense" [7]. This suggests that the pattern or texture of dense tissue should be considered while assessing mammograms. In addition, some texture features have been suggested to predict *BRCA1/2* carrier status, in contrast with PD [8].

A growing body of literature has considered mammographic texture features and their relationship with breast cancer risk. A recent review paper identified 17 original research articles [9]. These included early work by Manduca, et al. [6], who identified texture features based on the grey-level co-occurrence matrix (GLCM) of neighbouring pixels. Kontos, et al. [10] looked at a range of texture features with the aim to see how they are associated with PD based on both digital mammography and digital breast tomosynthesis (DBT). With limited sample size, they identified GLCM features (homogeneity, contrast and energy) were associated with PD from DBT and, to a lesser extent, digital mammography. Haberle, et al. [11] considered five types of texture features, finding that statistical features based on GLCM were strongly predictive of breast cancer, and found that PD did not add information to risk once texture features had been accounted for. Li, et al. [3] found that textural features predict breast cancer slightly better than semi-automated percent density. Keller, et al. [12] compared risk prediction models using PD along with texture features based on the GLCM, statistical moments, and run-length [13] and reported that texture features outperformed PD. However, there is not a great deal of consistency in textural features identified between studies, so more work in this area is critically important. Nielsen, et al. [14] developed a mammographic texture resemblance (MTR) marker based on multi-scale Gaussian features, which was found to have similar prediction performance compared with PD and could further improve the predictive ability when PD and MTR is combined.

While several studies have identified texture features for cancer prediction, many have been based on digitised film [9]. With the introduction of FFDM breast screening, there is a need to assess how best to assess the risk from textural features using digital mammograms. This is important partly because the properties of FFDM images differ from those of digitised films. For example, FFDM have a higher dynamic range than digitised film images [15] resulting in richer grey-level profiles; they also have different noise properties since the inherent granularity of screen-film mammography is not present in FFDM [16].

Very few studies have looked at texture features of original raw FFDM images. An additional issue with digital processed images is that one has to rely on manufacturers' proprietary processing algorithms before feature extraction, which may result in images from different machines less comparable. A recent review [9] on texture features for breast cancer risk found just two case-control studies based on raw FFDM including those by Chen, et al. [17] and Zheng, et al. [18]. There were just 156 cases in these studies (combined), and the case mammograms were from the

contralateral (unaffected) breast at breast cancer diagnosis. Thus, overall information on the ability of textural features to predict risk of breast cancer in this context is limited.

The aim of this paper is to develop a fully automated texture feature extraction system for raw digital mammograms, and to assess the prediction ability of textural features to stratify risk beyond volumetric percent density. *Fully automated* in the context of this paper refers to a texture feature extraction system, including any pre-processing procedure, which operates without any human intervention.

2 Methods

2.1 Setting and study design

Two case-control studies were designed using women recruited to the Predicting Risk Of breast Cancer At Screening (PROCAS) cohort, in Manchester, UK [19].

The first case-control study was for feature selection (the training set), and cases were women with cancer detected at first screen on entry to PROCAS. Women were matched approximately 3:1 (controls vs cases) by age, BMI, hormone replacement therapy (HRT) use and menopausal status. For feature selection the craniocaudal (CC) views of the contralateral breast for cases and the left breast for controls were used [20]. Unaffected breast were followed up and recorded and cases with bilateral cancer were excluded. The average follow-up time is 3.86 years for cases and 4.86 years for controls.

The second case-control study was used to validate the risk score (the validation set). Each woman had a normal screening mammogram (no cancer detected) on entry to PROCAS, but an interval or screen-detected cancer arose subsequently. The mammograms were approximately three years prior to diagnosis of breast cancer and were sampled independently from the same cohort as the training set. There is a small overlap of controls between the two datasets (n=45) representing 2.7% of the total number of controls in both datasets. Again women were matched approximately 3:1 (controls vs cases) by age, BMI, HRT use, menopausal status, as well as on year of mammogram at entry. Since the validation was done in a double-blind fashion, case-control status was unknown before validation, so a pre-defined list of side of breast for each woman was provided so that the contralateral breast (also CC views) for cases and the same side for controls were used. As with the first study, women with bilateral cancer were excluded. The average follow-up time to date of diagnosis is 3.03 years for cases and the average follow-up time is 4.28 years for controls.

2.2 Mammograms

All digital raw (“for processing”) mammograms were acquired using a GE Senographe system. The resolution of the mammograms was 10.628 pixels per mm. Percent volumetric density was assessed using Volpara 1.5.0 (Volpara Health Technologies, Wellington, New Zealand).

2.3 Texture features

Texture features were extracted from the whole breast as a single region after windowing. Specifically, the minimum pixel value (whitest area) in the breast region was used as the lower bound of the window, and the value at the 75th percentile of the pixel value range (darker areas) within the breast was taken as the upper bound. The lower and upper bounds of the window were then reset (lower bound to 1 and pixels on or above the upper bounds to 0, which as a result also inverted the image) and the rest of the pixel values were linearly rescaled between 0 and 1.

We generally follow the literature to decide whether a feature is to be considered. Statistical moments of pixel values from the windowed images were calculated directly in addition to the following secondary features: grey-level co-occurrence matrix (GLCM), neighbourhood grey-tone difference matrix (NGTDM), form and shape of breast boundary, run-length, and grey-level size zone matrix (GLSZM) [3, 6, 10, 11, 13, 20-23].

Texture features were extracted from images at their original resolution. In addition, since some features (GLCM, NGTDM, run-length, and GLSZM) are resolution sensitive and might be associated with risk differently at different scales, they were extracted at reduced resolutions, by factors of 2, 4, 6, 8, 16, 32, and 64 using bicubic interpolation [6].

All texture features were calculated using Matlab (Mathworks, Natick, MA). The Matlab package developed by Vallieres, et al. [24] was employed for computing the GLCM, NGTDM, run-length, and GLSZM features; and for these features, pixels were grouped equally into ten grey levels in forming the relevant matrices before computing the texture features. A total number of 327 features were identified to be investigated. The full list of texture features considered and their types, downsize factors, univariate goodness-of-fit statistics using the training set are provided in the supplementary file [see Additional file 1].

2.4 Statistical analysis

2.4.1 Feature selection and model building

An initial screening was performed to remove features whose absolute Pearson correlation with any other feature was greater than 0.95, where the feature taken forward was randomly selected. This resulted in a total of 112 candidate texture features.

Feature selection was based on the least absolute shrinkage and selection operator (LASSO) method, adjusted for age, BMI and volumetric PD. The tuning parameter that controls the extent of coefficient shrinkage was chosen by cross-validation. The final calibrated model was based on the one standard error rule, where the most parsimonious model whose error (deviance in this case) was within one standard error of the model with minimum cross-validation error (leave one out) was selected [25]. LASSO feature selection was performed using the implementation by Friedman, et al. [26] in R [27]. A single risk score based on the LASSO fit was taken forward for validation. In addition, Volpara Density Grade (VGD), a categorical version of estimated volumetric PD was also tested to see whether VGD adds information to volumetric PD or selected texture features.

2.4.2 Validation of risk score and components

The composite risk score as well as individual texture features identified by LASSO were validated in a two-stage double-blind fashion. A statistical analysis plan was drafted detailing the procedure of data exchange and statistical analysis. After identification of a limited set of textural features and a risk score to investigate further using the training data, CW calculated these features using anonymised mammograms from the validation set, and blind to case-control status. EH ran the initial statistical analysis for these features using the validation set, and then unblinded CW. The predictive ability of the risk score beyond volumetric PD was tested using conditional logistic regression. Deviance (or likelihood-ratio χ^2) and the matched concordance index (mC) [28] were calculated to test and measure prediction performance. Deviance is a likelihood based statistic, and in the context of logistic regression is equivalent to the sum of squared residuals. For model comparison, it is common practice to examine the change in deviance (likelihood-ratio χ^2) to measure relative model performance. mC is a modification of the concordance index (or area under the receiving operator characteristic, AUC) to matched case-control studies, and gives an average concordance index within matched groups. Some other features that were not selected by LASSO but had previously been identified to be important, and were observed to be univariately significant in the training set (i.e. standard deviation, coarseness and contrast as shown below), were also assessed in the validation case-control study. Since biologic phenotypes between screen-detected and interval cancers are different, the effects of texture features or volumetric PD on risk may also differ. To explore this, a series of multivariate models have been fitted with risk factors that are statistically significant in the univariate models, and additional interaction term between the image feature and indicator for screen-detected or interval cancer.

3 Results

3.1 Study characteristics

The training case-control study had a total of 264 cases and 787 controls, of which 199 cases were invasive, 63 were ductal carcinoma in situ (DCIS), and two unknown. The validation case-control study had a total of 317 cases and 931 controls, of which 277 were invasive, 39 were DCIS and one was unknown. The demographic characteristics of the women in the two studies are summarised in Table 1 which shows that age, BMI, and HRT use were well matched between cases and controls in both studies. As expected, volumetric PD was generally higher among cases than controls in both studies. The median 10-year Tyrer-Cuzick score was also higher for cases than controls in both studies. A majority of women had never used HRT and the percentage was slightly higher in the training set (60% for controls and 65% for cases in training set, vs. 51% for controls and 52% for cases in validation set; the differences between training and validation sets are significant with p-values of 0.0002 and 0.0019 respectively). In both studies around three quarters of women were postmenopausal, and the majority of women were ethnically white.

Table 1 is about here

3.2 Texture feature risk score development

Three features were selected from the training set using LASSO (the value of the LASSO tuning parameter=0.0402) and taken forward for validation in a combined risk score. They were all the GLCM feature *Sum Average* but calculated using images downsized by factors of 16, 32, and 64. *Sum Average* is a feature considered to capture the relation between clear and dense areas in an image (i.e. radiolucent and radiopaque) [29]. Table 2Table 2 shows the correlation coefficients between the three Sum Average features, volumetric PD, age, BMI, and other important features identified in the literature including standard deviation (SD), contrast (based on NGTDM), and coarseness calculated at the original image resolution. Coarseness measures the amount of local grey-level variation and contrast measures the amount of difference among all grey levels as well as the amount of local variation in grey level presented in the image [21]. SD is a histogram based feature so does not take into account spatial relationships between pixels.

Table 2 is about here

The *Sum Average* features at different resolutions were relatively highly and positively correlated (Spearman 0.74 to 0.88). There were weaker and negative associations between *Sum Average* features and age (-0.23 to -0.18) or BMI (-0.35 to -0.23). Volumetric PD was quite strongly and positively correlated with the *Sum Average* features (0.54 to 0.63).

Table 3 is about here

Table 3 (a) shows prediction performance of volumetric PD, three *Sum Average* features univariately, as well as additional texture features that have previously been identified in the literature and were significant in the training data, and taken forward to be assessed in the validation set as secondary measures.

In the training sample, all three *Sum Average* features outperformed the other univariate features in terms of χ^2, as well as achieving better mC than PD except SumAverage64. *Sum Average* downsized by a factor of 32 achieved the best result in terms of both χ^2 and mC (0.61). The performance of PD, SD, and contrast was similar, while coarseness was the least predictive in terms of χ^2. We have also tested Volpara Density Grade (VGD), a categorical version of estimated volumetric PD, finding it has a very similar predictive performance compared to volumetric PD (χ^2=20.19, degrees of freedom=3). A series of likelihood-ratio tests have shown that VGD does not add further information to either volumetric PD ($\Delta\chi^2$=3.36, p=0.3), or LASSO selected texture features such as SumAverage16 ($\Delta\chi^2$= 3.40, p=0.3).

The risk score taken forward for validation is a weighted linear combination of the three *Sum Average* features. The standardized weights (i.e. using z-scores where predictors were rescaled by their means and standard deviations before entering the model) are:

Risk score= 0.044*SumAverage16 + 0.036 *SumAverage32 + 0.066 *SumAverage64

where the means of the three features were respectively 0.0555, 0.0559 and 0.0566; the standard deviations were respectively 0.000238, 0.000430 and 0.000775. It can be seen that SumAverage64 contributed most to the score (0.066 / (0.066+0.036+0.044) = 45%). The risk score had a similar mC (0.60) to its *Sum Average* components.

Figure 1 is about here

Figure 1 shows the mC and its confidence intervals for the *Sum Average* features calculated at different resolutions, including those not selected by the LASSO algorithm. Generally mC increased as images were downsized up to a factor of 32, and was approximately flat at downsizing factors between 16 and 128.

Figure 2Figure 2 ***is about here***

To better understand the feature *Sum Average* and risk score, and see how the feature looks visually, example images with low and high values of risk scores but similar volumetric PDs in the training study are presented in Figure 2.

3.3 Validation of texture risk score

The regression results using the validation dataset in Table 3 (b) confirmed the predictive power of the texture risk score found in the training dataset. The standardized odds ratio was 1.36 (95% CI 1.20-1.55) with mC 0.58 (95% CI 0.54-0.62), which was broadly comparable with the development analysis using training set (mC=0.60). The risk score also achieved a better performance than volumetric PD in terms of deviance ($\Delta\chi^2$=10.55), indicating some evidence of preference of risk score relative to the PD [30] (logarithm PD $\Delta\chi^2$=6.95), as well as a similar mC (0.58 compared with 0.57 for PD). A series of likelihood-ratio tests showed that the risk score also added independent predictive information to volumetric PD ($\Delta\chi^2$ = 14.38, p = 0.0008), as well as Tyrer-Cuzick risk (logarithm transformed, $\Delta\chi^2$ = 22.43, p < 0.0001) and PD and Tyrer-Cuzick combined ($\Delta\chi^2$ = 10.22, p=0.001). On the other hand, once the risk score was taken into account, PD added little information ($\Delta\chi^2$ = 0.21, p= 0.7).

Looking at individual texture features, only the three *Sum Average* features and contrast were statistically significant. *Sum Average* based features also achieved the best fit in terms of deviance/χ^2 compared with other texture features and PD. Additionally, only the *Sum Average* based features added information to PD (the χ^2 test statistics were 5.16, 6.46, and 12.56 for *Sum Average* using images downsized by factors of 16, 32, and 64 respectively). This confirms that *Sum Average* at low resolutions is an independent risk factor. Other texture features did not add further information once the risk score was taken into account.

Modelling results showing the difference between screen-detected and interval cancers for statistically significant features are presented in Table 4. As Table 4 shows, with the exception of contrast, the difference in screen-detected and interval cancers is statistically significant; and texture features and volumetric PD have higher odds ratio for interval than screen-detected cancers.

Table 4 is about here

4 Discussion

This paper aimed to predict breast cancer risk with various texture features from raw digital mammograms. To achieve this, relevant features were extracted and the LASSO model was employed for feature selection. The risk score was validated using a separate set of cases and controls within the cohort.

The original raw mammogram files were pre-processed using a windowing technique. This effectively means that the darkest 25% pixels within the breast (mostly the uncompressed region) were set to be background. This is similar to the method used by Heine, et al. [20] for computing standard deviation. They eroded a 25% area from the edge of the breast in scanned film images, since they reported that the region in question could potentially interfere with further feature extraction. The breast edge contains the darkest pixels. In addition to standardising pixel intensities, another benefit of windowing is that image contrast is enhanced, making image appearance similar to that of film mammograms.

The texture features tested included many of those identified in previous studies, such as standard deviation of the pixel intensity values, NGTDM contrast, coarseness, and GLCM features. We also assessed some novel features that have been less well studied in the literature, including GLSZM based features that measure zonal effects and some form-based features such as diameter of a circle with the same area as the breast region.

The GLCM feature *Sum Average* at lower image resolutions was selected by LASSO in the training study. Based on its mathematical formulation (see Appendix) and visual assessment of some mammograms, one can show that this feature tends to identify dispersed patterns of density on a mammogram. It was slightly surprising that PD and some previously reported texture features such as standard deviation, contrast and coarseness were not selected, although contrast was significant and negatively associated with risk in both training and validation studies, in-line with Huo, et al. [21]. Other texture features such as standard deviation and coarseness however were not significant in the validation study. While it is interesting that only 3 features were selected out of 112 features by the LASSO algorithm, it is worth noting that the features that were not selected by LASSO are not necessarily unpredictive of risk. For example, the feature contrast was shown as predictive in both training and validation studies. Volumetric PD was not selected by LASSO either. This may be an indication that once some features were used, other features may no longer add information. This is supported by the likelihood-ratio test result that shows once risk score is taken into account, volumetric PD adds little information ($p= 0.7$). In the validation study, the three *Sum Average* based features achieved the best results among univariate predictors in terms of both deviance and mC. The risk score, a weighted combination of three *Sum Average* based features, has only obtained similar deviance or mC to its components univariately, suggesting *Sum Average* measured at one image resolution might be adequate. Although the *Sum Average* feature has been employed in some previous studies, it has not been identified as the strongest texture feature

previously. The reason for different findings might be due to differences in the methods used to compute textural features. Indeed, it is often difficult to determine precisely how a feature was computed in prior publications and so we have been careful to provide a detailed description of the *Sum Average* feature used here in the appendix. A lesser factor for differences might be different feature selection methods. Previous studies have often used stepwise regression for feature selection [6, 8, 11, 12, 21]. However, as pointed out by Hastie, et al. [25], stepwise regression often leads to poor results compared to a less greedy method such as LASSO.

We explored the risk score by visual inspection of mammograms. Those in Figure 2 are deliberately extreme, but they were chosen to show readers a clear demonstration that mammograms with a high risk score have more dispersed areas of bright pixels; whilst those with a low risk score do not. The example shows that a higher risk score helps to identify more widely dispersed dense patterns. In other words, it might capture an element of dense area that is (implicitly) not necessarily taken into account by volumetric density. As observed in Table 2, there is fairly high correlation between texture features and PD, so some of the effects of PD may be captured by texture features. Considering texture features improve prediction beyond PD, it is possible some spatial patterns of dense tissue may be related to risk in addition to relative amount of density. This interpretation also follows the mathematical formula for the feature. Downsizing is important because it enables the measure of spatial relationships between pixels at a greater distance, and so better measure wider areas of density. In summary, this feature seems to capture the distribution of dense tissue and our results suggest that mammograms with greater areas of high density are associated with higher risk.

Differences in prediction performance at different resolutions are due to change in patterns for each feature at those resolutions. Some texture features are more consistent than others when image were re-scaled. For example, the spearman correlation coefficient for Sum Average between downsize factors of 1 and 64 is -0.12, indicating weak association; while the spearman correlation coefficient for coarseness between downsize factors of 1 and 64 is 0.78, showing strong correlation. This means some factors such as coarseness are more consistent than others when image were re-scaled. Texture features such as those based on GLCM, typically measure spatial relationships between a pixel and its neighbouring pixels. As images were downsized, the neighbouring pixels become more distant, thus result in changes in feature patterns. Some features, such as coarseness, are relatively robust to such change in neighbourhood definition, while some features changed dramatically. This suggests that it is important to consider the impact of image resolution while analysing a certain texture feature. The implication is that a feature that predicts well at a given resolution may not perform well at another resolution. It is thus important to indicate the image resolution when exploring the prediction performance of a feature. This finding has also been observed elsewhere. For example, Haberle, et al. [11] reported that a GLCM feature based on the same set of mammograms but at different resolutions have either different (opposite) associations with PD or different associations with cancer risk. Manduca, et al. [6] also found texture features tended to predict risk better when they were extracted at reduced image

resolutions. For instance, the area under the receiver operating characteristic curve (AUC) of a feature increases from 0.50 to 0.60 when the images were downsized by factors of 2, 4, 8, and 16.

One contribution of our paper is that it shows how to extract a useful textural feature in a fully automated way from digital raw mammograms. Traditionally studies utilising image texture features for cancer prediction were based on scanned films e.g. [6, 21]. Also, there is concern that results from processed (i.e. for presentation) mammograms may not be generalizable since different manufacturers have their own proprietary processing algorithms, making the resulting images and their features potentially not fully comparable between different manufacturers and machines. This paper addresses the above concerns by using the raw FFDM, and has shown which texture features might be important for predicting breast cancer risk, and how the risk model can be improved by downsizing the images. It is anticipated that the method proposed in the paper would better facilitate breast cancer risk prediction by using digital mammograms.

There are several possible ways to expand our study. For example, our image pre-processing method did not consider acquisition parameters, such as compression force, and thickness of the compressed breast and breast edge. It is possible that employing these acquisition parameters may lead to better image pre-processing and ultimately risk prediction. Another direction is to externally validate the method on a different population with different characteristics such as ethnicity and parity. In particular, we note that more than 92% of our study population were white, and more than 88% were parous. The use of larger and diverse datasets would allow for additional breast cancer risk factors to be adjusted in the model. Also, the mammograms used in our analysis were all acquired from a GE system. It would be interesting to test our method on mammograms produced by other brands of machines. For transferability of our method, digital mammograms from other machines may be re-scaled to the same resolution as in this paper before feature extraction. There is also potential that our method can be adapted for digitised films. It would also be interesting to compare our method to recent advancement in deep learning [31], which employs unsupervised machine learning to detect useful image features. Finally, this study focuses on the CC view of mammograms. It is possible that texture features that are predictive for cancer risk may be different for mediolateral oblique (MLO) view mammograms. The issue with using the MLO view of mammograms is how to treat the pectoral muscle. One possible approach is to remove the pectoral muscle before feature extraction. This requires an automated pectoral muscle removal algorithm (e.g. [17]) since our ultimate aim is to develop a fully automated risk prediction system. The additional information from MLO views may assist to better predict breast cancer risk than using CC views alone.

5 Conclusion

This paper has shown that texture features are useful for predicting breast cancer risk using raw digital mammograms. Important texture features previously identified in the literature as well as some novel features were tested. The feature selection method LASSO was adopted to finalise the feature set taken forward for validation.

Among various features tested including standard deviation, coarseness, contrast and volumetric PD, we found the GLCM feature *Sum Average* at low image resolution was the strongest predictor of breast cancer risk, and added independent information to volumetric PD. An image standardisation method was adopted to pre-process the digital raw mammograms before feature extraction, making it likely that our approach would have merit on other mammogram machines. However, while the selected features and calibrated model were internally validated in a separate case-control study with consistent results, our findings and risk algorithm would benefit from further studies to externally validate them.

6 List of abbreviations

- AUC: area under the receiver operating characteristic curve
- BI-RADS: Breast Imaging Reporting and Data System
- BMI: body mass index
- CC: craniocaudal
- CI: confidence interval
- DBT: digital breast tomosynthesis
- DCIS: ductal carcinoma in situ
- GLCM: grey-level co-occurrence matrix
- GLSZM: grey-level size zone matrix
- HRT: hormone replacement therapy
- LASSO: least absolute shrinkage and selection operator
- mC: matched concordance index
- MLO: mediolateral oblique
- MTR: Mammographic texture resemblance
- NGTDM: neighbourhood grey-tone difference matrix
- OR: odds ratio
- PD: percent density
- PROCAS: Predicting Risk Of breast Cancer At Screening
- SD: standard deviation
- VGD: Volpara Density Grade

7 Declarations

7.1 Ethics approval and consent to participate

The PROCAS study was approved by Central Manchester Research Ethics Committee (reference: 09/H1008/81) and the consent was obtained from study participants at the time of screening.

7.2 Consent for publication

Not applicable.

7.3 Availability of data and material

The datasets used for the current study are available upon reasonable request from the corresponding author.

7.4 Competing interests

The authors declare that they have no competing interests.

7.5 Funding

This research is partially funded by the Cancer Research UK (grant number C569/A16891). This work was supported by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research programme (reference number RP-PG-0707-10031: "Improvement in risk prediction, early detection and prevention of breast cancer") and the Genesis Prevention Appeal (references GA10-033 and GA13-006). The views expressed are those of the author(s) and not necessarily those of the Cancer Research UK, NHS, the NIHR, or the Department of Health.

7.6 Authors' contributions

CW developed the algorithm to extract the texture features, performed the statistical analysis including feature selection, interpreted the results, and drafted the manuscript. AB made substantial contribution to texture feature development, advice on statistical analysis and interpretation, and helped to draft the manuscript. EH made substantial contribution to data distribution and model validation, and helped to draft the manuscript. DGE conceived the PROCAS study. JC and SA conceived the study, interpreted the results, and help to draft the manuscript. All authors read and approved the final manuscript.

7.7 Acknowledgements

The authors would like to thank the women who agreed to take part in the PROCAS study and other members of the PROCAS group including the study radiologists and advanced radiographic practitioners, and study staff for recruitment and data collection.

8 Appendix

Sum Average is a statistical texture feature computed from grey-level co-occurrence matrix (GLCM) constructed by considering how often pairs of pixels with specific values and in a specified spatial relationship occur in an image. Sum Average is defined as [32]:

$$\left[\sum_{i,j}(i+j)\cdot p(i,j)\right]/(2I^2).$$

where *I* is the total number of grey levels, and $p(i,j)$ denotes the probability occurrence of a pixel at grey level *i* is in a defined spatial relationship to a pixel at grey level *j* in an image. Such probability can be created by firstly counting the frequencies of pairs of pixels at different grey levels with a defined spatial relationship (in this study, all eight directions surrounding a pixel were counted), and then normalized so that the sum of the elements of the GLCM is equal to 1. I^2 is included in the formula so as to make this feature comparable between different sizes of GLCMs.

Also in this study, the number of grey levels of 10 was adopted (i.e. *I*=10), and the pixels within the breast region was divided into 10 levels in such a way that each level has equal probability. We tested using number of grey levels other than 10 and the feature pattern changed only marginally – for instance we tested using 5 grey levels and the correlation coefficient with Sum Average using 10 grey levels was 0.97.

Sum Average can be seen as the weighted (by grey levels) sum of GLCM elements. Thus this texture is likely to have higher value if many high grey level pixels are clustered in a blob.

In addition to Sum Average, the following GLCM features were tested:

- Contrast: $\sum_{i,j}|i-j|^2 p(i,j)$
- Correlation: $\sum_{i,j}\frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i\sigma_j}$, where μ, σ are means and standard deviations of corresponding rows and columns of GLCM.
- Dissimilarity: $\sum_{i,j}|i-j|p(i,j)$
- Energy: $\sum_{i,j} p(i,j)^2$
- Entropy: $\sum_{i,j} -p(i,j)\log\big(p(i,j)\big)$
- Homogeneity: $\sum_{i,j}\frac{p(i,j)}{1+|i-j|}$
- Variance: $\left[\sum_{i,j}\left((i-\mu_i)^2+\left(j-\mu_j\right)^2\right)\cdot p(i,j)\right]/(2I^2)$

Similar to GLCM, a *neighbourhood grey-tone difference matrix* (NGTDM) could be constructed, and relevant texture features could be extracted by computing the summary statistics of NGTDM. The NGTDM is a vector (column matrix) constructed by firstly calculating the average grey-tone over a neighbourhood centred at, but excluding (k,l):

$$\bar{A}_\iota = \bar{A}(k,l) = \frac{1}{W-1}\left[\sum_{m=-d}^{d}\sum_{n=-d}^{d} f(k+m,l+n)\right]$$

where $(m,n) \neq (0,0)$ (i.e. excluding the (*k,l*)); f(*k,l*) is the grey tone of any pixel at (*k,l*) having grey tone value *i*; *d* specifies the neighbourhood size (d=1 in this case); and $W = (2d+1)^2$. Then the ith element of the NGTDM is:

$$s(i) = \begin{cases} \sum |i-\bar{A}_\iota|, & for\ i \in N_i, if\ N_i \neq 0 \\ 0, & otherwise \end{cases}$$

where N_i is the set of all pixels having grey tone *i* (exluding the peripheral regions of width *d*).

Coarseness is defined as $[\epsilon + \sum_i p_i s(i)]^{-1}$, where ε is a small number (2^{-52} in this case) to prevent it becoming infinite; *p* is the probability of occurrence of the corresponding intensity value.

Contrast is defined as $\left[\frac{1}{N_g(N_g-1)}\sum_i\sum_j p_i p_j (i-j)^2\right]\left[\frac{1}{n^2}\sum_i s(i)\right]$, where N_g is the total number of different grey levels in the image; n=N-2d.

Coarseness and contrast have been successfully applied for classification of cancer or BRCA1/2 status in the literature, such as Huo, et al. [21] and Kontos, et al. [10] where these features were described. As with GLCM, the pixels within the breast region were equally divided into 10 levels. In addition to coarseness and contrast, the following NGTDM features were tested:

- Busyness: $[\sum_i p_i s(i)]/\left[\sum_{i,j} ip_i - jp_j\right], p_i \neq 0, p_j \neq 0$
- Complexity: $\sum_{i,j}\left\{|i-j|/\left(n^2(p_i+p_j)\right)\right\}\{p_i s(i) + p_j s(j)\}, p_i \neq 0, p_j \neq 0$
- Strength: $\left[\sum_{i,j}(p_i+p_j)(i-j)^2\right]/[\epsilon + \sum_i s(i)], p_i \neq 0, p_j \neq 0$

Similar to GLCM and NGTDM, run-length features were extracted from the run-length matrix. Let $p(i,j)$ be the number of runs with pixels of grey level *i* and run-length *j*, n_r be the total number of runs, and n_p be the number of pixels in the region of interest:

- Short run emphasis (SRE): $\frac{1}{n_r}\sum_i \sum_j \frac{p(i,j)}{j^2}$
- Long run emphasis (LRE): $\frac{1}{n_r}\sum_i \sum_j p(i,j) \cdot j^2$
- Grey-Level Nonuniformity (GLN): $\frac{1}{n_r}\sum_i (\sum_j p(i,j))^2$
- Run Length Nonuniformity (RLN): $\frac{1}{n_r}\sum_j (\sum_i p(i,j))^2$
- Run Percentage (RP): n_r/n_p
- Low Grey-Level Run Emphasis (LGRE): $\frac{1}{n_r}\sum_i \sum_j \frac{p(i,j)}{i^2}$
- High Grey-Level Run Emphasis (HGRE): $\frac{1}{n_r}\sum_i \sum_j p(i,j) \cdot i^2$
- Short Run Low Grey-Level Emphasis (SRLGE): $\frac{1}{n_r}\sum_i \sum_j \frac{p(i,j)}{i^2 \cdot j^2}$
- Short Run High Grey-Level Emphasis (SRHGE): $\frac{1}{n_r}\sum_i \sum_j \frac{p(i,j)\cdot i^2}{j^2}$
- Long Run Low Grey-Level Emphasis (LRLGE): $\frac{1}{n_r}\sum_i \sum_j \frac{p(i,j)\cdot j^2}{i^2}$
- Long Run High Grey-Level Emphasis (LRHGE): $\frac{1}{n_r}\sum_i \sum_j p(i,j) \cdot i^2 \cdot j^2$
- Grey-Level Variance (GLV): $\sqrt{\frac{1}{I \cdot J}\sum_i \sum_j [p(i,j) \cdot i - \mu]^2}$, where $\mu = \frac{1}{I \cdot J}\sum_i \sum_j p(i,j) \cdot i$.
- Run-Length Variance (RLV): $\sqrt{\frac{1}{I \cdot J}\sum_i \sum_j [p(i,j) \cdot j - \mu]^2}$, where $\mu = \frac{1}{I \cdot J}\sum_i \sum_j p(i,j) \cdot j$.

The GLSZM features are similar to run-length features but with focus on sizes of zones instead of collinear pixels (i.e. runs). In GLSZM, $p(i,j)$ is defined as the number of zones with pixels of grey level *i* and area *j*, and the same formulas for run-length features can be used to compute GLSZM features.

As for shape-based features, convex area measures the number of pixels in the smallest convex polygon that contains the breast; equivalent diameter measures the diameter of a circle with the same area as the breast; extent measures the ratio of pixels in the breast to pixels in the total bounding box; major axis length is the length of the longest diameter of an ellipse that has the

same normalized second central moments as the breast region; similarly minor axis length is the length of the minor axis of an ellipse that has the same normalized second central moments as the breast; and solidity is the ratio of breast area and its convex area.

A software implementing the method in this paper has been made available for Windows operating system (https://doi.org/10.6084/m9.figshare.4994429.v2). Upon launching the software, a dialog box would prompt asking users which mammogram file(s) to examine and where the results to be saved. The software would then compute the texture features without further user input and saved the results in a spreadsheet at the location the user specified.

Differences from density assessment case-control studies:

The number of cases and controls differs from a report (submitted elsewhere) that compared density methods using the same women. The reasons are as follows.

Firstly, the training case-control study was a subset of one with 317 cases and 952 controls. Three women were excluded due to linkage errors between mammograms and questionnaire data (although they could be subsequently incorporated we decided to present the training data as it was undertaken before validation). We also excluded women with unknown BMI and volumetric PD at the time of analysis (79 and 37 women with missing BMI and PD respectively). An additional 100 controls were removed during conditional logistic regressions because they did not have matched cases as a result of the above exclusions.

The validation case-control study originally had 338 cases and 1014 controls. 23 women were excluded because of unavailability of mammograms at the time of validation (either no mammograms provided for some women at the given side; or only MLO views were available but no CCs). 64 women were further removed because the side of cancer (left or right) was unknown; and two women were further excluded due to lack of volumetric PD data at the time of validation. A further 15 controls were removed during conditional logistic regressions because they had no matched cases as a result of the above exclusions.

9 References

[1] V. Assi, J. Warwick, J. Cuzick, and S. W. Duffy, "Clinical and epidemiological issues in mammographic density," *Nature Reviews Clinical Oncology,* vol. 9, pp. 33-40, Jan 2012.

[2] N. F. Boyd, L. J. Martin, L. M. Sun, H. Guo, A. Chiarelli, G. Hislop*, et al.*, "Body size, mammographic density, and breast cancer risk," *Cancer Epidemiology Biomarkers & Prevention,* vol. 15, pp. 2086-2092, Nov 2006.

[3] J. Li, L. Szekely, L. Eriksson, B. Heddson, A. Sundbom, K. Czene*, et al.*, "High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer," *Breast Cancer Research,* vol. 14, p. R114, 2012.

[4] B. M. Keller, D. L. Nathan, Y. Wang, Y. J. Zheng, J. C. Gee, E. F. Conant*, et al.*, "Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation," *Medical Physics,* vol. 39, pp. 4903-4917, Aug 2012.

[5] J. N. Wolfe, "Breast Patterns as an Index of Risk for Developing Breast-Cancer," *American Journal of Roentgenology,* vol. 126, pp. 1130-1139, 1976.
[6] A. Manduca, M. J. Carston, J. J. Heine, C. G. Scott, V. S. Pankratz, K. R. Brandt*, et al.*, "Texture Features from Mammographic Images and Risk of Breast Cancer," *Cancer Epidemiology Biomarkers & Prevention,* vol. 18, pp. 837-845, Mar 2009.
[7] R. American College of and B.-R. Committee, *ACR BI-RADS atlas : breast imaging reporting and data system*. Reston, VA: American College of Radiology, 2013.
[8] G. L. Gierach, H. Li, J. T. Loud, M. H. Greene, C. K. Chow, L. Lan*, et al.*, "Relationships between computer-extracted mammographic texture pattern features and BRCA1/2 mutation status: a cross-sectional study," *Breast Cancer Research,* vol. 16, pp. 1-16, 2014.
[9] A. Gastounioti, E. F. Conant, and D. Kontos, "Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment," *Breast Cancer Research,* vol. 18, Sep 20 2016.
[10] D. Kontos, L. C. Ikejimba, P. R. Bakic, A. B. Troxel, E. F. Conant, and A. D. A. Maidment, "Analysis of Parenchymal Texture with Digital Breast Tomosynthesis: Comparison with Digital Mammography and Implications for Cancer Risk Assessment," *Radiology,* vol. 261, pp. 80-91, Oct 2011.
[11] L. Haberle, F. Wagner, P. A. Fasching, S. M. Jud, K. Heusinger, C. R. Loehberg*, et al.*, "Characterizing mammographic images by using generic texture features," *Breast Cancer Research,* vol. 14, 2012.
[12] B. M. Keller, J. B. Chen, E. F. Conant, and D. Kontos, "Breast density and parenchymal texture measures as potential risk factors for Estrogen-Receptor positive breast cancer," *Medical Imaging 2014: Computer-Aided Diagnosis,* vol. 9035, 2014.
[13] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing,* vol. 4, pp. 172-179, 1975/06/01 1975.
[14] M. Nielsen, C. M. Vachon, C. G. Scott, K. Chernoff, G. Karemore, N. Karssemeijer*, et al.*, "Mammographic texture resemblance generalizes as an independent risk factor for breast cancer," *Breast Cancer Research,* vol. 16, 2014.
[15] S. Suryanarayanan, A. Karellas, S. Vedantham, H. Ved, S. P. Baker, and C. J. D'Orsi, "Flat-panel digital mammography system: Contrast-detail comparison between screen-film radiographs and hard-copy images," *Radiology,* vol. 225, pp. 801-807, Dec 2002.
[16] M. J. Yaffe, "Basic Physics of Digital Mammography," in *Digital Mammography*, U. Bick and F. Diekmann, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 1-11.
[17] X. Chen, E. Moschidis, C. Taylor, and S. Astley, "Breast Cancer Risk Analysis Based on a Novel Segmentation Framework for Digital Mammograms," *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2014, Pt I,* vol. 8673, pp. 536-543, 2014.
[18] Y. J. Zheng, B. M. Keller, S. Ray, Y. Wang, E. F. Conant, J. C. Gee*, et al.*, "Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment," *Medical Physics,* vol. 42, pp. 4149-4160, Jul 2015.
[19] D. G. R. Evans, J. Warwick, S. M. Astley, P. Stavrinos, S. Sahin, S. Ingham*, et al.*, "Assessing Individual Breast Cancer Risk within the U.K. National Health Service Breast Screening Program: A New Paradigm for Cancer Prevention," *Cancer Prevention Research,* vol. 5, pp. 943-951, 2012.
[20] J. J. Heine, C. G. Scott, T. A. Sellers, K. R. Brandt, D. J. Serie, F. F. Wu*, et al.*, "A Novel Automated Mammographic Density Measure and Breast Cancer Risk," *Journal of the National Cancer Institute,* vol. 104, pp. 1028-1037, Jul 2012.

[21] Z. Huo, M. L. Giger, O. I. Olopade, D. E. Wolverton, B. L. Weber, C. E. Metz*, et al.*, "Computerized Analysis of Digitized Mammograms of BRCA1 and BRCA2 Gene Mutation Carriers," *Radiology,* vol. 225, pp. 519-526, 2002.

[22] M. Amadasun and R. King, "Textural features corresponding to textural properties," *Systems, Man and Cybernetics, IEEE Transactions on,* vol. 19, pp. 1264-1274, 1989.

[23] A. Chu, C. M. Sehgal, and J. F. Greenleaf, "Use of Gray Value Distribution of Run Lengths for Texture Analysis," *Pattern Recognition Letters,* vol. 11, pp. 415-419, Jun 1990.

[24] M. Vallieres, C. R. Freeman, S. R. Skamene, and I. El Naqa, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," *Physics in Medicine and Biology,* vol. 60, pp. 5471-5496, Jul 21 2015.

[25] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning : data mining, inference, and prediction*, 2nd ed. New York, NY: Springer, 2009.

[26] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *2010,* vol. 33, p. 22, 2010-02-02 2010.

[27] R Core Team. (2016). *R: A language and environment for statistical computing*. Available: https://www.R-project.org/

[28] A. R. Brentnall, J. Cuzick, J. Field, and S. W. Duffy, "A concordance index for matched case-control studies with applications in cancer risk," *Statistics in Medicine,* vol. 34, pp. 396-405, Feb 10 2015.

[29] E. S. Gadelmawla, A. E. Eladawi, B. Abouelatta, and I. M. Elewa, "Investigation of the cutting conditions in milling operations using image texture features," *Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture,* vol. 222, pp. 1395-1404, Nov 2008.

[30] K. P. Burnham, D. R. Anderson, and K. P. Huyvaert, "AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons," *Behavioral Ecology and Sociobiology,* vol. 65, pp. 23-35, Jan 2011.

[31] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. F. Diao, C. Igel*, et al.*, "Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring," *Ieee Transactions on Medical Imaging,* vol. 35, pp. 1322-1331, May 2016.

[32] D. Assefa, H. Keller, C. Menard, N. Laperriere, R. J. Ferrari, and I. Yeung, "Robust texture features for response monitoring of glioblastoma multiforme on T1-weighted and T2-FLAIR MR images: A preliminary investigation in terms of identification and segmentation," *Medical Physics,* vol. 37, pp. 1722-1736, Apr 2010.

10 Figures

Figure 1 Matched concordance index (mC) for Sum Average at different image downsize factors, with bootstrap 95% confidence intervals (CI)

Figure 2 Comparison of mammograms (for presentation purpose processed images were shown) with two of the lowest (a) and highest (b) standardized risk scores. All mammograms have similar Volumetric PDs around 10%

Figure 2(a) Mammograms with low risk scores (-1.7 and -1.3 respectively). Volumetric PDs are 10.1% and 10.2 respectively.

Figure 2(b) Mammograms with high risk scores (3.2 and 2.0 respectively). Volumetric PDs are 9.9% and 10.0% respectively.

11 Additional files

File name: Additional file 1

- Word document format (.docx)
- Title: Univariate modelling results from training dataset
- This table shows the univariate modelling results of all candidate texture features considered using training dataset

Table 1 Demographics of training set (cancers detected at first screen on entry to the PROCAS study) and validation set (cancers detected at a subsequent screen or between screening rounds)

	Training set			*Validation set*		
	Controls	*Cases*	*p-*	*Controls*	*Cases*	*p-*
	N (%)	*N (%)*	*value*	*N (%)*	*N (%)*	*value*
Age at consent (y)			0.9994			0.9997
<50	44 (6)	16 (6)		46 (5)	16 (5)	
50-54	200 (25)	65 (25)		193 (21)	64 (20)	
55-59	150 (19)	51 (19)		164 (18)	58 (18)	
60-64	172 (22)	57 (22)		286 (31)	96 (30)	
65-69	166 (21)	57 (22)		195 (21)	67 (21)	
70+	55 (7)	18 (7)		47 (5)	16 (5)	
HRT use			0.2234			0.9646
Unknown	11 (1)	7 (3)		22 (2)	6 (2)	
Never	473 (60)	171 (65)		475 (51)	165 (52)	
Previous	262 (33)	72 (27)		329 (35)	110 (35)	
Current	41 (5)	14 (5)		105 (11)	36 (11)	
BMI (kg/m^2)			0.9797			0.9408
Unknown	-	-		-	1 (0)	
<25	241 (31)	80 (30)		335 (36)	117 (37)	
25-29	289 (37)	96 (36)		341 (37)	113 (36)	
≥30	257 (33)	88 (33)		255 (27)	86 (27)	
Menopausal status			0.9914			0.9887
Unknown	16 (2)	7 (3)		32 (3)	12 (4)	
Perimenopausal	94 (12)	32 (12)		134 (14)	46 (15)	
Postmenopausal	591 (75)	196 (74)		698 (75)	237 (75)	
Premenopausal	86 (11)	29 (11)		67 (7)	22 (7)	
Ethnic origin			0.0411			0.2229
Other/unknown	38 (5)	22 (8)		81 (9)	35 (11)	
White	749 (95)	242 (92)		850 (91)	282 (89)	
Parity			0.8143			0.0351
Unknown	-	-		1 (0)	4 (1)	
Nulliparous	97 (12)	34 (13)		90 (10)	44 (14)	
Parous	690 (88)	230 (87)		840 (90)	269 (85)	

	Training set				*Validation set*			
	Controls		*Cases*	*p-*	*Controls*		*Cases*	*p-*
	N (%)		*N (%)*	*value*	*N (%)*		*N (%)*	*value*
Tyrer-Cuzick (10y risk-%): *(median, Q1-Q3)*	2.73	(2.19 - 3.60)	2.82 (2.29 - 3.88)	0.0028	2.68	(2.09 - 3.55)	2.91 (2.24 - 4.03)	<.0001
Volumetric PD: *(median, Q1-Q3)*	5.34	(4.06 - 7.35)	5.88 (4.62 - 8.55)	0.0003	4.73	(3.50 - 6.92)	5.31 (3.79 - 7.57)	0.0041

HRT: hormone replacement therapy; Q1: 25th percentile; Q3: 75th percentile; p-values, from likelihood-ratio chi-square tests, indicate whether there are significant difference between cases and controls.

Table 2 Spearman correlation coefficients between age, BMI, PD and texture features

	Age	BMI	Volumetric PD	Sum Average 16	Sum Average 32	Sum Average 64	SD	Coarseness	Contrast
Age	1								
BMI	0.03	1							
Volumetric PD	-0.14	-0.57	1						
Sum Average 16	-0.19	-0.35	0.63	1					
Sum Average 32	-0.23	-0.33	0.63	0.81	1				
Sum Average 64	-0.18	-0.23	0.54	0.74	0.88	1			
SD	-0.16	-0.19	0.46	0.27	0.32	0.31	1		
Coarseness	-0.12	-0.62	0.79	0.49	0.58	0.51	0.53	1	
Contrast	0.15	0.34	-0.74	-0.45	-0.52	-0.52	-0.64	-0.80	1

PD: percent density; BMI: body mass index; SD: standard deviation; Sum Average 16, 32, 64: texture feature Sum Average using images downsized by a factor of 16, 32, and 64.

Table 3 Univariate modelling results from training and validation datasets

Table 3 (a) Univariate modelling results from training dataset

Parameter	*Standardized Odds Ratio*	*95% CI for Odds Ratio*	χ^2	*p-value*	*mC*	*95% CI for mC*
Coarseness	1.22	(1.06 - 1.41)	7.28	6.98E-03	0.58	(0.53 - 0.62)
Contrast	0.73	(0.62 - 0.85)	16.75	4.27E-05	0.40	(0.36 - 0.45)
SD	1.32	(1.13 - 1.54)	13.11	2.94E-04	0.57	(0.52 - 0.61)
SumAverage16	1.52	(1.31 - 1.77)	31.26	2.25E-08	0.61	(0.56 - 0.65)
SumAverage32	1.52	(1.31 - 1.77)	31.75	1.75E-08	0.61	(0.56 - 0.66)
SumAverage64	1.48	(1.28 - 1.71)	29.07	6.98E-08	0.58	(0.53 - 0.63)
Volumetric PD	1.36	(1.18 - 1.57)	18.05	2.16E-05	0.59	(0.55 - 0.64)

Total number of observations (N) = 1051, including 264 cases and 787 controls.

Table 3 (b) Univariate modelling results using validation dataset

Parameter	*Standardized Odds Ratio*	*95% CI for Odds Ratio*	χ^2	*p-value*	*mC*	*95% CI for mC*
Risk score	1.36	(1.20 - 1.55)	22.39	2.22E-06	0.58	(0.54 - 0.62)
Coarseness	1.06	(0.92 - 1.22)	0.61	4.34E-01	0.50	(0.46 - 0.54)
Contrast	0.87	(0.76 - 0.99)	4.33	3.75E-02	0.46	(0.42 - 0.50)
SD	1.01	(0.89 - 1.15)	0.03	8.55E-01	0.50	(0.45 - 0.54)
SumAverage16	1.29	(1.14 - 1.47)	15.37	8.85E-05	0.58	(0.54 - 0.62)
SumAverage32	1.32	(1.16 - 1.50)	17.55	2.80E-05	0.58	(0.53 - 0.62)
SumAverage64	1.38	(1.21 - 1.57)	23.81	1.06E-06	0.59	(0.55 - 0.63)
Volumetric PD	1.27	(1.11 - 1.46)	11.84	5.80E-04	0.57	(0.53 - 0.61)

Total number of observations (N) = 1248, including 317 cases and 931 controls.
Standardized odds ratio: change in odds for a standard deviation (in controls) increase in predictors, adjusted for age and BMI; CI: confidence interval; mC: matched concordance index; SD: standard deviation; SumAverage 16, 32, 64: texture feature Sum Average using images downsized by a factor of 16, 32, and 64. PD: percent density.

Table 4 Modelling results for screen-detected and interval cancer

	Risk score	*Contrast*	*SumAverage16*	*SumAverage32*	*SumAverage64*	*Volumetric PD*
Standardized OR for screen-detected cancer	1.15 (0.99,1.35)	0.88 (0.75,1.04)	1.09 (0.93,1.26)	1.11 (0.95,1.30)	1.20 (1.02,1.40)	1.13 (0.94,1.35)
Standardized OR for interval cancer	2.09 (1.59 - 2.74)	0.84 (0.66 - 1.06)	2.12 (1.59 - 2.81)	2.15 (1.62 - 2.86)	1.91 (1.48 - 2.47)	1.53 (1.21 - 1.92)
$\Delta\chi^2$	15.27	0.15	18.53	17.84	9.80	4.38
p-value	0.0001	0.70	<0.0001	<0.0001	0.002	0.036

Standardized OR (odds ratio): change in odds for a standard deviation (in controls) increase in image features; their 95% confidence intervals in brackets; SumAverage 16, 32, 64: texture feature Sum Average using images downsized by a factor of 16, 32, and 64. PD: percent density. $\Delta\chi^2$ and p-values refer to likelihood-ratio tests on whether there is significant difference between screen-detected and interval cancers (i.e. significance of interaction terms).