Radiology

ORIGINAL RESEARCH • EVIDENCE-BASED PRACTICE

Prediction of Upstaging in Ductal Carcinoma in Situ Based on Mammographic Radiomic Features

Rui Hou, PhD • Lars J. Grimm, MD, MHS • Maciej A. Mazurowski, PhD • Jeffrey R. Marks, PhD • Lorraine M. King, PhD • Carlo C. Maley, PhD • Thomas Lynch, PhD • Marja van Oirsouw, BSc • Keith Rogers, PhD • Nicholas Stone, PhD • Matthew Wallis, MD • Jonas Teuwen, PhD • Jelle Wesseling, MD, PhD • E. Shelley Hwang, MD, MPH • Joseph Y. Lo, PhD

From the Departments of Radiology (R.H., L.J.G., J.Y.L.) and Surgery (J.R.M., L.M.K., T.L., E.S.H.), Duke University Medical Center, Box 3513, Durham, NC 27710; Department of Electrical and Computer Engineering, Pratt School of Engineering, Duke University, Durham, NC (R.H.); School of Life Sciences, Arizona State University, Tempe, Ariz (C.C.M.); Cranfield Forensic Institute, Cranfield University, Cranfield, UK (K.R.); School of Physics and Astronomy, College of Engineering, Mathematics and Physical Sciences, Physics Building, Streatham Campus, University of Exeter, Exeter, UK (N.S.); Cambridge Breast Unit and NIHR Cambridge Biomedical Research Center, Cambridge University Hospitals NHS Trust, Cambridge Biomedical Campus, Cambridge, UK (M.W.); and Netherlands Cancer Institute, Amsterdam, the Netherlands (M.v.O., J.T., J.W.). Received March 1, 2021; revision requested April 23; revision received September 24; accepted October 19. **Address correspondence to** E.S.H. (e-mail: *shelley.hwang@duke.edu*).

Study supported in part by the National Cancer Institute of the National Institutes of Health (U01-CA214183, R01-CA185138), U.S. Department of Defense Breast Cancer Research Program (W81XWH-14-1-0473), Breast Cancer Research Foundation (BCRF-16-183, BCRF-17-073), Cancer Research UK and Dutch Cancer Society (C38317/A24043), and an equipment grant from Nvidia.

C.C.M. supported by National Institutes of Health (grants U54 CA217376, P01 CA91955, R01 CA185138, and R01 CA140657). M.W. supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014).

Conflicts of interest are listed at the end of this article.

Radiology 2022; 000:1–9 • https://doi.org/10.1148/radiol.210407 • Content codes: BR IN

Background: Improving diagnosis of ductal carcinoma in situ (DCIS) before surgery is important in choosing optimal patient management strategies. However, patients may harbor occult invasive disease not detected until definitive surgery.

Purpose: To assess the performance and clinical utility of mammographic radiomic features in the prediction of occult invasive cancer among women diagnosed with DCIS on the basis of core biopsy findings.

Materials and Methods: In this Health Insurance Portability and Accountability Act–compliant retrospective study, digital magnification mammographic images were collected from women who underwent breast core-needle biopsy for calcifications that was performed at a single institution between September 2008 and April 2017 and yielded a diagnosis of DCIS. The database query was directed at asymptomatic women with calcifications without a mass, architectural distortion, asymmetric density, or palpable disease. Logistic regression with regularization was used. Differences across training and internal test set by upstaging rate, age, lesion size, and estrogen and progesterone receptor status were assessed by using the Kruskal-Wallis or χ^2 test.

Results: The study consisted of 700 women with DCIS (age range, 40–89 years; mean age, 59 years ± 10 [standard deviation]), including 114 with lesions (16.3%) upstaged to invasive cancer at subsequent surgery. The sample was split randomly into 400 women for the training set and 300 for the testing set (mean ages: training set, 59 years ± 10; test set, 59 years ± 10; $P = .85$). A total of 109 radiomic and four clinical features were extracted. The best model on the test set by using all radiomic and clinical features helped predict upstaging with an area under the receiver operating characteristic curve of 0.71 (95% CI: 0.62, 0.79). For a fixed high sensitivity (90%), the model yielded a specificity of 22%, a negative predictive value of 92%, and an odds ratio of 2.4 (95% CI: 1.8, 3.2). High specificity (90%) corresponded to a sensitivity of 37%, positive predictive value of 41%, and odds ratio of 5.0 (95% CI: 2.8, 9.0).

Conclusion: Machine learning models that use radiomic features applied to mammographic calcifications may help predict upstaging of ductal carcinoma in situ, which can refine clinical decision making and treatment planning.

Ductal carcinoma in situ (DCIS), a stage 0 form of breast cancer, accounts for about 14.9% of all new breast cancer diagnoses (1). Although DCIS is not life threatening and many cases of DCIS may never progress to invasive cancer (2,3), it is a potential precursor to invasive ductal carcinoma (4). Among patients with core-needle biopsy–proven DCIS, concurrent invasive ductal carcinoma is found at the time of definitive surgery (upstaging) in 0%–26% of women (5–10). Currently, it is standard of care for DCIS to be excised; thus, occult invasive disease, if present, is detected at surgical excision. However, nonsurgical management strategies, including active surveillance, are being explored to address concerns about DCIS overtreatment. A major challenge to the feasibility of active surveillance is delayed detection of occult invasive cancer previously detected at surgery. Therefore, improving the presurgical diagnosis of occult invasive cancer in women with newly diagnosed DCIS can have important clinical implications and can assist providers and patients in choosing optimal management strategies (11,12).

Previous studies have investigated the factors associated with DCIS upstaging to invasive cancer in patients with DCIS at core-needle biopsy. In a large meta-analysis, Brennan et al (10) found that mammographic abnormalities including mass, architectural distortion, asymmetry, lesion palpability, and lesion size were significantly associated with upstaging, but DCIS manifesting as pure calcifications can also harbor occult invasive disease. The task of predicting

Abbreviations

AUC = area under receiver operating characteristic curve, DCIS = ductal carcinoma in situ

Summary

Mammographic radiomic features may help predict occult invasive disease in core-needle biopsy–proven ductal carcinoma in situ.

Key Results

- Mammograms from 700 women with asymptomatic ductal carcinoma in situ (DCIS) were used to develop logistic regression with regularization models to predict upstaging. Upstaging rate was 16.3% (114 of 700) and a total of 109 radiomic and four clinical features were extracted.
- Combining clinical and radiomic features provided the best prediction performance (area under receiver operating characteristic curve, 0.71).
- The model can hypothetically provide 90% sensitivity and 92% negative predictive value for guiding patients with DCIS who are considering active surveillance.

upstaging in women with DCIS who present only with calcifications has been persistently difficult for radiologists (13).

By using digitally extracted mammographic radiomic features, we conducted a retrospective cohort study to investigate whether radiomics with a machine learning approach could be used to distinguish pure DCIS from DCIS with occult invasive cancer. Specifically, we focused on women with DCIS who presented only with calcifications because these women may be eligible for de-escalation treatment strategies such as active surveillance. In addition, we evaluated the use of our model in surgical treatment planning for patients undergoing surgery. The purpose of our study was to assess the performance and clinical utility of mammographic radiomic features in the prediction of occult invasive cancer among women diagnosed with DCIS on the basis of core biopsy findings.

Materials and Methods

Study Sample

We searched medical records for all patients who underwent mammography, had calcifications, and were diagnosed with DCIS on the basis of a 9-gauge vacuum-assisted core-needle biopsy at a single health system between September 2008 and April 2017. We excluded women who presented with a mass, asymmetry, architectural distortion, or palpable disease; were younger than 40 years; had synchronous contralateral breast cancer; or had a history of breast cancer or surgery. Estrogen receptor and progesterone receptor status, as well as nuclear grade, were recorded from the initial core-needle biopsy reports. The subsequent surgical excision pathologic reports were also reviewed to determine whether there was subsequent invasive cancer (upstaging). All initial pathologic interpretations were made by pathologists with specialty training in breast surgical pathology; no new interpretations were made for this study.

Training and test sets were split randomly within each class to balance for upstaging rate. After the split, we confirmed that the ages of women in the training and test sets were not significantly different. Among the training set, 140 patients (105 with pure DCIS and 35 with upstaged DCIS) were reported in a previous study (14). This previous study focused on analyzing the adjunctive roles of two classes (atypical ductal hyperplasia and invasive ductal carcinoma) in improving prediction performance.

The prebiopsy digital mammograms were collected. All women underwent digital mammography with magnification views acquired by systems with a magnification factor of either 1.5× or 1.8× (Senographe Essential, GE Healthcare; or Selenia Dimensions, Hologic).

The institutional review board approved this retrospective study and waived written informed consent. The study complies with the Health Insurance Portability and Accountability Act.

Feature Extraction

Each breast lesion was identified by a fellowship-trained breast radiologist (L.J.G., with 6 years of experience), who was provided all available images and reports to guide his annotations. Calcifications were automatically segmented by using a U-Net convolutional neural network trained on a previously reported computer vision algorithm (14,15).

Model Building and Testing

The pipeline of this study is depicted in Figure 1. Before models were built, all features were standardized with zero mean and unit variance separately for each vendor (GE or Hologic), with training and test set normalized separately. The data set was randomly shuffled and was divided into training and internal test sets by balancing their upstaging rate. The training set was used for repeated cross validation. The internal test set was reserved and not used until the final testing. *Negative* corresponded to pure DCIS without upstaging and *positive* indicated DCIS that was upstaged to invasive cancer at the time of surgery. This presented a difficult image classification challenge because these two classes were both diagnosed as having only DCIS at initial core biopsy. For the subset of women with positive results later relabeled as upstaged after surgery, each lesion therefore was composed of a mix of both DCIS and invasive components in close proximity. Thus, any image features describing DCIS versus invasive disease would also be mixed together within the same lesion. We further analyzed the performance of radiomic or clinical features separately and in combination, and with or without feature selection. Because there were only four clinical features, feature selection was not applied on the model with only clinical features. The differences among all five models are illustrated in Table 1.

During training, fivefold cross-validations were repeated 200 times after random shuffling of the training set, which reduced bias from sample ordering across the training folds. During each repeat, we use a nested cross-validation; the outer loop handles resampling while the inner loop performs hyperparameter tuning and stabilized feature selection (16,17) using default parameters of the gridsearchCV function (Python module scikit-learn 0.20). The five models were trained by using logistic regression with L2 regularization (18).

For the test set evaluation, a final classifier was created by fixing the hyperparameters and features to the most frequently selected values during cross-validation training,

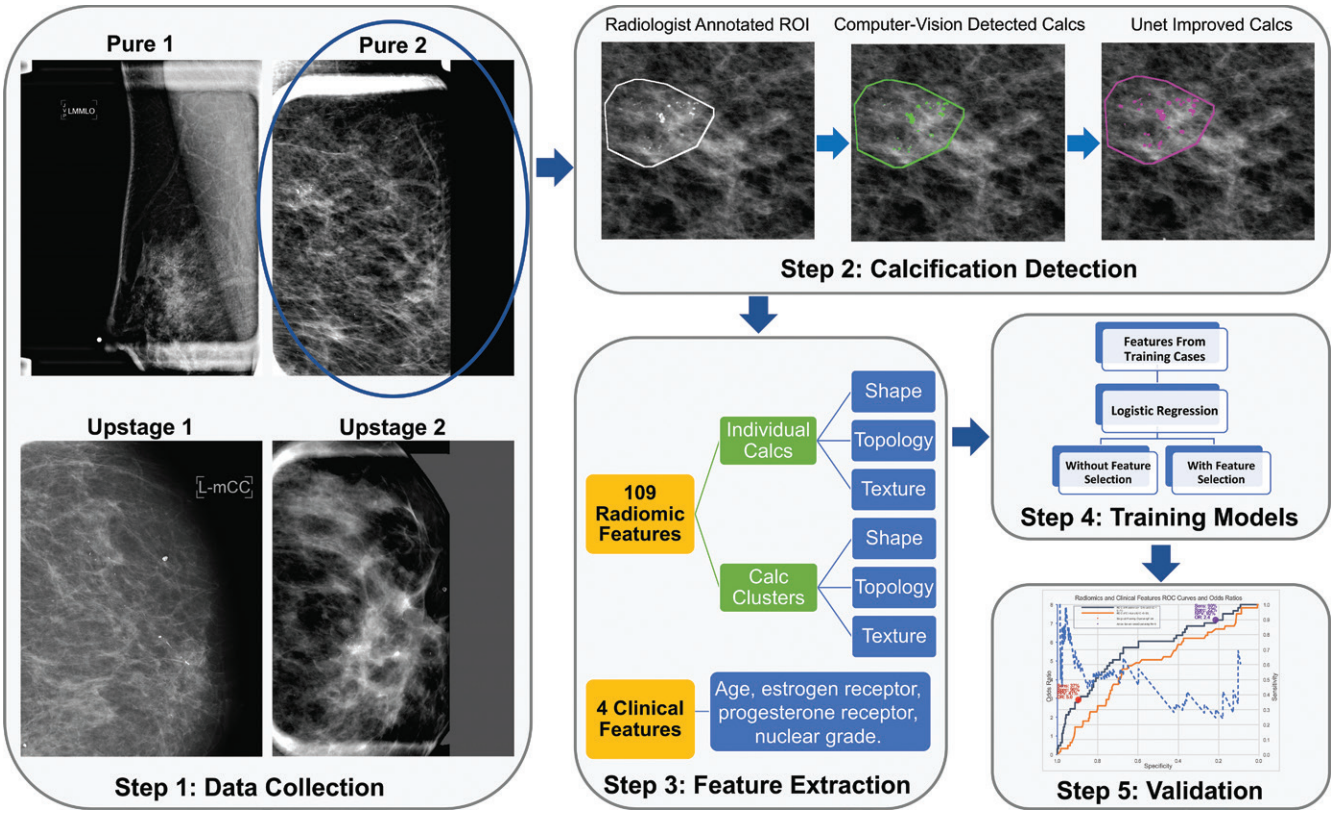

Figure 1: Illustration of the study pipeline (step 1). A total of 700 women were identified (step 2). Lesion annotations were masked by a breast radiologist, and calcifications (calcs) were masked by a computer vision–based algorithm and a deep learning–based U-net segmentation network (step 3). A total of 109 radiomic features and four clinical features were collected (step 4). Models with those extracted features and training data were trained (step 5). Selected models were validated on test data. NPV = negative predictive value, OR = odds ratio, PPV = positive predictive value, ROC = receiver operating characteristic, ROI = region of interest, sens = sensitivity, spec = specificity.

which were applied over the entire training set to create the fixed model for testing.

Statistical Analysis

Performance was assessed by the area under the receiver operating characteristic curve (AUC) with 95% CIs (19). Characteristics of the study sample and markers were summarized according to individual diagnosis at core biopsy. Differences between DCIS training and test sets were statistically analyzed with Kruskal-Wallis tests for continuous variables and χ^2 test for categorical variables. All statistical tests were two-tailed, with a significance level of .05. Feature extraction and machine learning models were implemented in Python. Statistical analysis was implemented in R software (R Project for Statistical Computing). The entire code is publicly available on GitLab (*https://gitlab.oit.duke.edu/railabs/LoGroup/mammographic-radiomics-to-predict-dcis-upstaging*).

Table 1: Comparison of Model Attributes and Performance during Validation and Testing

Model Name	Radiomic Features	Clinical Features	Feature Selection	Validation AUC	Test AUC
1	Yes	...	...	0.63 (0.56, 0.71)	...
2	Yes	...	Yes	0.68 (0.61, 0.75)	0.69 (0.60, 0.77)
3	...	Yes	...	0.59 (0.51, 0.67)	0.60 (0.51, 0.69)
4	Yes	Yes	...	0.63 (0.56, 0.71)	0.71 (0.62, 0.79)
5	Yes	Yes	Yes	0.68 (0.61, 0.75)	...

Note.—Data in parentheses are 95% CIs. AUC = area under the receiver operating characteristic curve.

Results

Study Sample Characteristics

We retrospectively identified all DCIS with stereotactic biopsied calcifications (n = 1791) from January 1, 2008, through May 1, 2017. The study excluded patients with a mass, architectural distortion, asymmetric density, or palpable disease (n = 491), patients younger than 40 years (n = 72), and patients with any type of previous cancers or breast surgery (n = 528). The final study sample included 700 consecutive women. A flowchart of patients included for analysis is shown in Figure 2. There were 114 of 700 women (16.3%) with DCIS upstaged to invasive cancer. A total of 400 women (335 with pure DCIS and 65 with upstaged DCIS; 16.3% upstage rate) were randomly selected for the training set. The remaining 300 women (251 with pure DCIS and 49 with upstaged DCIS; 16.3% upstage rate) were

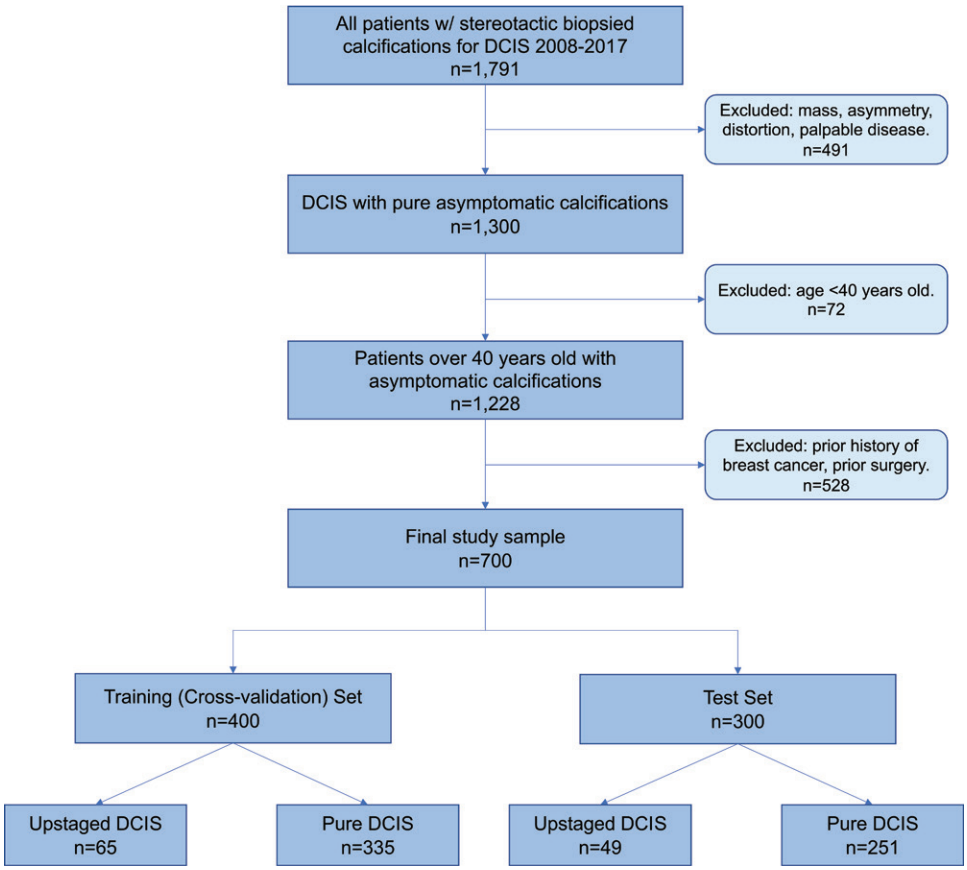

Figure 2: Study inclusion flowchart of patients with ductal carcinoma in situ (DCIS).

Table 2: Demographic Characteristics of Women and Lesion Characteristics in Total and in Training and Test Sets

Characteristic	Training (n = 400)	Test (n = 300)	P Value
Upstaged			.98*
No. of upstaged DCIS cases	65	49	
Upstaging rate (%)	16.3	16.3	
Mean age at diagnosis (y)†	59 (40–89)	59 (40–88)	.85‡
Mean lesion size (mm)†	27.1 (1.4–174.2)	28.3 (1.8–184.6)	.08‡
Estrogen receptor status			.87*
Positive	319 (79.8)	245 (81.7)	
Negative	74 (18.5)	55 (18.3)	
Unknown	7 (1.7)	0 (0.0)	
Progesterone receptor status			.10*
Positive	287 (71.8)	209 (69.7)	
Negative	105 (26.2)	90 (30.0)	
Unknown	8 (2.0)	1 (0)	
Nuclear grade			.08*
Low (grade I)	21 (5.2)	6 (2.0)	
Intermediate (grade II)	149 (37.3)	102 (34.0)	
High (grade III)	230 (57.5)	192 (64.0)	

Note.—Unless otherwise indicated, data in parentheses are percentages. DCIS = ductal carcinoma in situ.

*χ^2 test.

† Data in parentheses are the range.

‡ Kruskal-Wallis test.

Table 3: Radiomic and Clinical Features

Feature	Description
Individual calcification-level features	
Shape	
Calcification perimeter	Length of calcification contour
Calcification area	No. of pixels inside calcification contour × pixel size
Calcification circularity	$\frac{4\pi}{(MC\ Perimeter)^2} * (MC\ Area)$; measure of "roundness"
Calcification eccentricity	Fit into an ellipse
Calcification major axis	Length of calcification major axis
Calcification minor axis	Length of calcification minor axis
Calcification Hu moments × 7	Hu moment invariants: image moment invariants regarding translation, scale, and rotation (22)
Topology	
Calcification distance to centroid	Distance to calcification cluster centroid
Calcification distance to closest	Distance to nearest calcification neighbor
Calcification normalized degree	Sum of normalized weights of calcification degree (21), number of edges incident to the calc
Texture	
Calcification background × 2	Mean and standard deviation of surrounding background pixel intensities
Calcification foreground × 2	Mean and standard deviation of calcification pixel intensities
Calcification GLCM × 4	Measures computed from gray level co-occurrence matrices
Calcification cluster-level features × 13	
Shape	
Cluster area	Area of cluster
Cluster major axis	Length of cluster's major axis
Cluster eccentricity	Eccentricity of cluster
Topology	
Cluster calc number	No. of calcifications in this cluster
Cluster coverage	$\frac{\Sigma\,(all\ individual\ calcifications'\ areas)}{cluster's\ area}$
Texture	
Background × 2	Mean and standard deviation of cluster's surrounding pixel intensities
Foreground × 2	Mean and standard deviation of calcification pixel intensities in cluster
GLCMs × 4	Computed from GLCMs for cluster
Diagnosis-level features × 4	
Clinical	
Nuclear grade	Low and intermediate vs high
Estrogen receptor	Negative versus positive
Progesterone receptor	Negative versus positive
Age	In years

Note.—Four statistical pooling including mean, standard deviation, minimum, and maximum were applied on each feature vector. GLCM = gray-level co-occurence matrix.

reserved for testing. The characteristics of women with DCIS and markers were well matched between women in the training and test sets, without significant differences between sets in age at presentation, DCIS extent, hormone receptor status, or nuclear grade. Characteristics of the study sample and markers with *P* values are in Table 2. A total of 109 radiomic features were collected, representing each lesion's imaging characteristics to describe both individual calcifications and the overall calcification cluster (20). In addition, four clinical features were extracted from the core biopsy pathologic report: estrogen receptor, progesterone receptor, nuclear grade, and age at diagnosis. Detailed descriptions of these features are shown in Table 3.

Training by Cross-Validation

The model performance obtained with use of clinical features only, either individually or in combination, was consistently poor. Each individual feature resulted in an AUC ranging from 0.51 to 0.57: age, 0.53 (95% CI: 0.45, 0.60); estrogen receptor, 0.56 (95% CI: 0.44, 0.68); progesterone receptor, 0.57 (95% CI: 0.45, 0.69); nuclear grade, 0.55 (95% CI: 0.46, 0.64). None of the 95% CIs of any model excluded chance (AUC, 0.5). The AUC for the model with all four clinical features was 0.59 (95% CI: 0.51, 0.67).

Performances for the models involving different combinations of radiomics and clinical features are shown in Table 1. The use of all radiomic features generated a training AUC of 0.63 (95% CI: 0.56, 0.71). Then, applying feature selection resulted in a set of 11 selected features and an improved AUC performance of 0.68 (95% CI: 0.61, 0.75). The selected features were standard deviation, maximum, minimum, and mean of individual calcifications normalized degree (21); standard deviation of calcifications mean background intensity; standard deviation and maximum of calcifications distance to cluster centroid; number of calcifications in the lesion; standard deviation of calcifications mean intensity; maximum of calcifications minor axis; and maximum of calcifications third Hu moment (22).

Combining both radiomic and clinical features achieved a similar AUC of 0.63 (95% CI: 0.56, 0.71) to the model with radiomic features alone (AUC, 0.63; 95% CI: 0.56, 0.71; $P = .85$). After feature selection, the highest-performing model was feature selection with both radiomic and clinical features (AUC, 0.68; 95% CI: 0.61, 0.75), which incorporated the same 11 radiomics features as the model of radiomic features along with feature selection. In other words, none of the added clinical variables were selected, and

therefore the models of feature selection on either radiomic features alone or both radiomic and clinical features were the same.

Independent Testing

On the basis of our training performance, a subset of models was selected for the final evaluation on the independent test set. The model of radiomic features without feature selection was excluded because of lower performance than that of radiomic features with feature selection. The model combining radiomic and clinical features with feature selection was excluded because it was exactly the same model (based on the same 11 features) as the model of radiomics features along with feature selection. This resulted in three final models, with AUC and 95% CIs shown in Table 1. The model that included both radiomic and clinical features demonstrated the highest performance (AUC, 0.71;

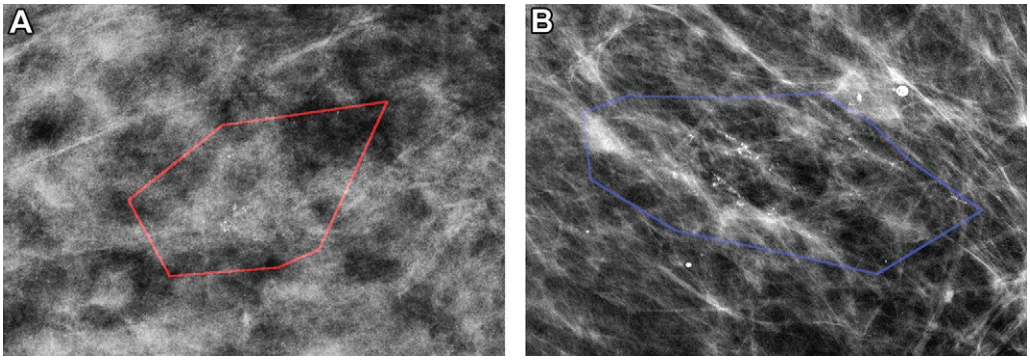

Figure 3: Mammographic images of patients with biopsy-proven ductal carcinoma in situ (DCIS). **(A)** A 55-year-old woman (right magnification craniocaudal view) diagnosed with DCIS only; model correctly classified as negative findings. **(B)** A 64-year-old woman (left magnification mediolateral oblique view) with DCIS at core biopsy but subsequently upstaged to invasive disease; model correctly classified as positive findings. Red and blue polygons show lesions annotated by the radiologist.

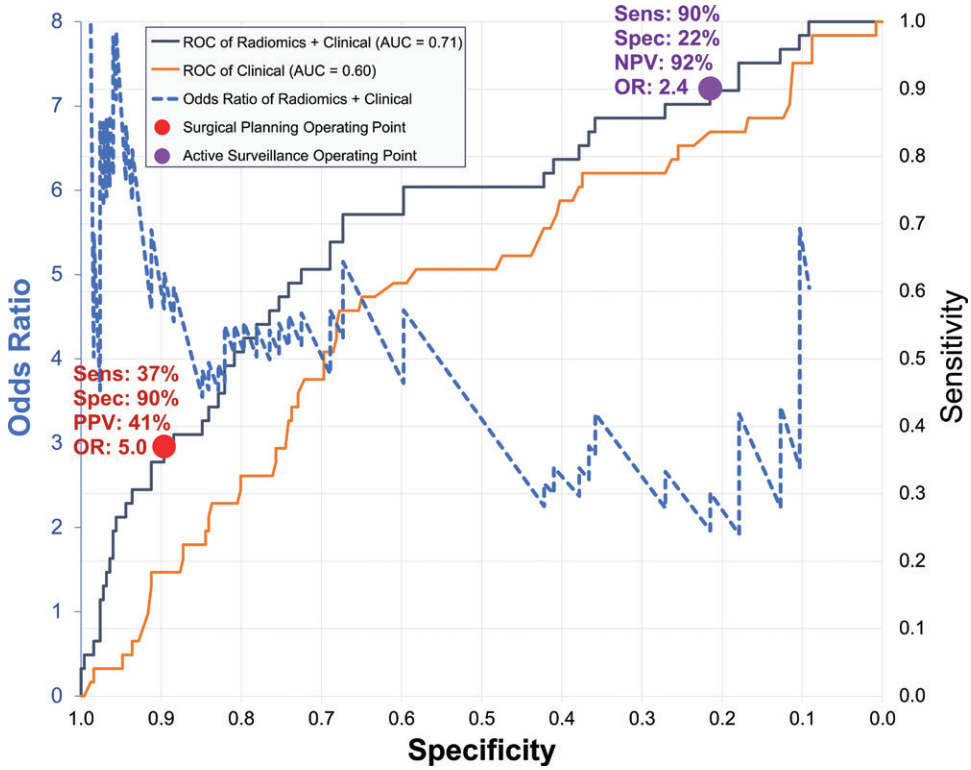

Figure 4: Graph shows receiver operating characteristic (ROC) curves and odds ratios (ORs) of prediction models. Receiver operating characteristic curves are shown for two models: one using radiomic and clinical features (gray) and one using clinical features alone (orange). Both receiver operating characteristic curves are plotted as sensitivity (secondary vertical axis on the right) versus specificity (horizontal axis). Blue dashed line is OR curve, plotted as OR (primary vertical axis on left) versus specificity (horizontal axis). Two operating points are shown with symbols and are described in the text: high-sensitivity active surveillance (purple circle); high-specificity surgical planning for sentinel node biopsy alongside with lesion removal surgery (red circle). AUC = area under receiver operating characteristic curve, NPV = negative predictive value, PPV = positive predictive value, Sens = sensitivity, Spec = specificity.

Table 4: Prediction Performance for Diagnostic Criteria on Test Set

Criteria	Upstaging Rate		Specificity	Sensitivity	Odds Ratio†
	High Risk*	Low Risk			
Model with active surveillance	18.3 (44/241)	84.7 (5/59)	22 (54/251)	90 (44/49)	2.4 (1.8, 3.2)
Model with surgical planning	40.9 (18/44)	12.1 (31/256)	90 (225/251)	37 (18/49)	5.0 (2.8, 9.0)
Estrogen receptor status	23.6 (13/55)	14.7 (36/245)	83 (209/251)	27 (13/49)	1.80 (0.88, 3.68)
Progesterone receptor status	24.4 (22/90)	12.4 (26/209)	73 (183/251)	46 (22/48)	2.28 (1.21, 4.29)
Nuclear grade	18.2 (35/192)	13.0 (14/108)	38 (94/251)	71 (35/49)	1.50 (0.77, 2.93)

Note.—Data are percentages; data in parentheses are numerator/denominator unless otherwise indicated. Table is based on test set of 300 patients (251 with pure and 49 with upstaged ductal carcinoma in situ). Progesterone receptor data are not available for one patient. "Model" refers to the same model using both radiomics and clinical features without feature selection, thresholded at different cutoff values for active surveillance versus surgical planning scenarios.

* Definition of high risk for each individual feature: estrogen receptor status, negative; progesterone receptor status, negative; nuclear grade, III.

† Data in parentheses are 95% CIs.

95% CI: 0.62, 0.79), but the other model of radiomic features along with feature selection performed similarly (AUC, 0.69; 95% CI: 0.660, 0.77). All testing results were slightly better than training while still remaining within the 95% CIs, reflecting inevitable consequences of data sampling, even for this relatively large cohort. Figure 3 shows correctly classified examples from both classes: pure DCIS and upstaged DCIS.

Clinical Management Strategies

To consider hypothetical clinical management strategies, Figure 4 applies the best model, combining radiomic and clinical features, on the test data to depict the relationship between odds ratio and the receiver operating characteristic curve. The odds ratio is greatest when the specificity is high and sensitivity is low, and it drops with decreasing specificity and increasing sensitivity. Therefore, depending on the clinical circumstances, we can select different, hypothetical operating points. The purple circle is a high-sensitivity operating point that may be appropriate for active surveillance of women with pure DCIS, where a fixed 90% sensitivity yields 22% specificity (95% CI: 17, 27), 92% negative predictive value (95% CI: 89, 93), and odds ratio of 2.4 (95% CI: 1.8, 3.2). Alternatively, the red circle depicts a high-specificity operating point for women with high-risk DCIS to consider combination surgery comprising standard breast-conserving surgery with concurrent sentinel lymph node biopsy, in which a fixed 90% specificity yields 37% sensitivity (95% CI: 24, 51), 41% positive predictive value (95% CI: 32, 49), and an odds ratio of 5.0 (95% CI: 2.8, 9.0).

We compared the performance of our best model to key clinical features associated with low-risk DCIS. The predictive performance of individual features of estrogen receptor, progesterone receptor, and nuclear grade are shown in Table 4. For the purposes of calculating sensitivity, specificity, and odds ratio, high-risk outcomes are considered as the positive class, and low-risk outcomes are considered as negative. The radiomics models had the highest performance to identify women with upstaged DCIS, compared with individual clinical features of risk.

Discussion

Although many ductal carcinomas in situ (DCIS) may not progress to invasive cancer, current standard of care requires that all patients with DCIS undergo surgical excision. Improving the presurgical diagnosis of upstaged DCIS is important to determine eligibility for alternative management strategies. Our study showed that radiomic features from mammography have the potential to predict upstaging of DCIS. The combination of clinical and radiomic features provided the best prediction performance (area under the receiver operating characteristic curve [AUC], 0.71; 95% CI: 0.62, 0.79). From that receiver operating characteristic curve, two hypothetical operating points were described: active surveillance of women not likely to have invasive cancer at 92% negative predictive value and odds ratio of 2.4, and inclusion of sentinel lymph node biopsy surgery for women likely to be upstaged at 41% positive predictive value and odds ratio of 5.0. Notably, this model did not perform significantly better than the model with radiomic features alone (P = .11). Both radiomics models performed substantially better than the clinical features alone that have been used for this task previously (ie, age, estrogen receptor/progesterone receptor status, and nuclear grade). None of the individual clinical features performed well enough to exclude chance, and none were chosen in the feature selection.

Related research evaluated radiologist performance for predicting DCIS upstaging in a two-stage observer study on a total of 300 women (13). Mean performance of nine radiologists increased to an AUC of 0.77 (95% CI: 0.62, 0.85; P = .045) after the development of a set of consensus criteria. Given the wide CIs, however, the radiologists' performance is similar to that in our study. Our most selected radiomics feature of "normalized degree" describes how calcifications extend over a lesion area, which is similar to the feature in the radiologists' consensus criteria that densely packed calcifications are related to upstaging.

Recent work in breast MRI radiomics and pathologic radiomics (23–25) suggests the potential for integrating markers from different domains to improve performance.

There are three notable differences between our study and previous reports. First, our study developed a mammography radiomics model to predict DCIS upstaging, whereas previous studies (5,10) were based on clinical, radiologic, and histologic findings. Second, we specifically excluded lesions with well-established clinical features associated with upstaging, such as mass, architectural distortion, palpability, and asymmetry (26,27). We excluded these lesions because they confer a well-established higher risk for upstaging, are readily identified by radiologists, and already serve as the basis for exclusion criteria in active surveillance trials. Therefore, constraining this study to lesions with only calcifications made the task of predicting upstaging much more difficult, but the results have greater clinical relevance by enabling changes in management. Third, our study uses the largest DCIS mammography data set to date, with training and testing sets that are fivefold greater than the size of the next largest study, which included only 140 women with DCIS (35 women with upstaged DCIS) (14). Notably, that study used 113 features, whereas we performed the same task using a subset of 11 radiomic features. When other DCIS data sets become available, the model should undergo external validation to confirm its generalizability.

Classification models for upstaging have the potential to inform clinical management of patients with DCIS. Active surveillance has recently been proposed as a potential management strategy for patients with low-risk DCIS and allows patients to forego surgery initially, opting instead for close monitoring, including regularly scheduled mammography. By intention, our study already excluded higher-risk women not suitable for active surveillance (ie, those with mass, asymmetry, architectural distortion, palpability, and previous cancer), which may otherwise have greatly increased the sensitivity. Although hypothetical and applied retrospectively, the 92% negative predictive value shows promise for guiding patients who are considering active surveillance. As a complementary alternative, women identified as having high-risk DCIS may undergo combination surgery comprising standard breast-conserving surgery with concurrent sentinel lymph node biopsy. For women with occult invasive disease, concurrent sentinel lymph node biopsy would obviate a second operation to assess nodal status, thus reducing morbidity and streamlining clinical management. The benefit of concurrent sentinel lymph node biopsy must be weighed carefully against the cost of false-positive results, however, as sentinel lymph node biopsy procedures may be associated with up to 5% morbidity, including lymphedema and paresthesias. By way of context, molecular predictors such as DCIS score and DCISionRT have hazard ratios of 1.97 and 2.03, respectively, in identifying cohorts at increased risk for recurrence following lumpectomy for DCIS (28,29).

Our study had limitations. First, despite considerable differences in training strategies, the performances of the radiomics models were similar (AUCs of approximately 0.71), which may indicate an upper bound in the use of radiomic features from mammography for this difficult diagnostic task. Specifically, the handcrafted radiomic features were on the basis of characteristics that radiologists deemed important. However, automatic feature extraction, such as features from deep learning models, may have the ability to capture additional differences between these classes (30). Second, although these performances may seem modest, the challenging task of predicting upstaging was made more difficult by excluding conspicuous features, such as a mass or distortion. Our best model already outperforms existing nonradiomic criteria, providing potentially clinically relevant performance in terms of sensitivity versus specificity and odds ratios. As the field of radiology seeks to refine patient selection for personalized and risk-based treatment pathways, radiomic approaches will become increasingly relevant, augmenting standard clinical data and routine radiologic review. Third, the sensitivity and specificity of the two operating points we described were estimated from the internal test receiver operating characteristic. For future external testing, cutoffs in model output values should be established during the training so that the performances reflect the true uncertainty of independent testing. Finally, although the data set was split into training versus test sets while matching three key factors (age, lesion size, and prevalence), there was still sampling bias that caused the test performance to be better than training. This sampling bias may be minimized by resampling, but that may affect both the training and testing performances differently across multiple models.

In conclusion, we found that radiomic features derived from mammography can classify occult invasive disease in ductal carcinoma in situ (DCIS), with performance superior to that of clinical criteria alone. This suggests potential to use imaging algorithms to improve patient care and assist in the selection of patients for clinical trials. Additional ongoing efforts in our team include the use of deep learning prediction models and acquiring external test data from other centers and imaging platforms. The use of such tools for risk stratification may provide a tractable way forward to enable safe de-escalation of treatment in low-risk clinical settings. These combined efforts are aimed to allow better risk stratification, thus enabling a tractable way forward toward risk-based treatment for DCIS.

Author contributions: Guarantors of integrity of entire study, **R.H., E.S.H., J.Y.L.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **R.H., L.J.G., M.v.O., E.S.H.**; clinical studies, **L.J.G., J.R.M., M.v.O., J.W., E.S.H.**; experimental studies, **R.H., L.J.G., M.A.M., M.v.O., J.Y.L.**; statistical analysis, **R.H., M.A.M., N.S., E.S.H.**; and manuscript editing, **R.H., L.J.G., J.R.M., L.M.K., C.C.M., T.L., M.v.O., K.R., N.S., M.W., J.T., J.W., E.S.H., J.Y.L.**

Disclosures of conflicts of interest: **R.H.** No relevant relationships. **L.J.G.** Grants from ECOG/Acrin (TMIST) Alliance for Clinical Trials in Oncology Foundation AUR Breast Cancer Research Foundation AD Anderson Cancer Center; consulting fees from Hologic; payment for lectures from Medscape Reference; payment for expert testimony from Hare, Wynn, Newell and Newton LLP, leadership or fiduciary role in Society of Breast Imaging. **M.A.M.** No relevant relationships. **J.R.M.** Grant from University of Utah Subaward; participation on a Data Safety Monitoring board or advisory board at Duke University. **L.M.K.** No relevant relationships. **C.C.M.**

Leadership or fiduciary role in International Society for Evolution, Ecology and Cancer. **T.L.** No relevant relationships. **M.v.O.** No relevant relationships. **K.R.** No relevant relationships. **N.S.** Participation on a Data Safety or advisory board at University of Exeter. **M.W.** No relevant relationships. **J.T.** No relevant relationships. **J.W.** Member Research Council KWF Dutch Cancer Society; Member Scientific Advisory Board Member Dutch Expert Center for Screening; Advisor for the population-based breast screening program by the National Institute for Public Health and the Environment on behalf of the Dutch Society of Pathology. **E.S.H.** Consulting fees from AstraZeneca; payment for lectures from Merck; participation on a Data Safety or advisory board at Duke University; leadership or fiduciary role from Immunis Clinetic; stock options from Clinetic. **J.Y.L.** Equipment grant from Nvidia.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020;70(1):7–30.
2. Ryser MD, Weaver DL, Zhao F, et al. Cancer outcomes in DCIS patients without locoregional treatment. J Natl Cancer Inst 2019;111(9):952–960.
3. Sanders ME, Schuyler PA, Dupont WD, Page DL. The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. Cancer 2005;103(12):2481–2484.
4. Hieken TJ, Krishnamurthy V, Farolan M, Velasco JM. Long-term outcome of DCIS patients: p53 as a biomarker of ipsilateral recurrence. J Clin Oncol 2011;29(27_suppl):39.
5. Schulz S, Sinn P, Golatta M, et al. Prediction of underestimated invasiveness in patients with ductal carcinoma in situ of the breast on percutaneous biopsy as rationale for recommending concurrent sentinel lymph node biopsy. Breast 2013;22(4):537–542.
6. Sim YT, Litherland J, Lindsay E, et al. Upgrade of ductal carcinoma in situ on core biopsies to invasive disease at final surgery: a retrospective review across the Scottish Breast Screening Programme. Clin Radiol 2015;70(5):502–506.
7. Chan MY, Lim S. Predictors of invasive breast cancer in ductal carcinoma in situ initially diagnosed by core biopsy. Asian J Surg 2010;33(2):76–82.
8. Mannu GS, Wang Z, Broggio J, et al. Invasive breast cancer and breast cancer mortality after ductal carcinoma in situ in women attending for breast screening in England, 1988-2014: population based observational cohort study. BMJ 2020;369:m1570.
9. Han JS, Molberg KH, Sarode V. Predictors of invasion and axillary lymph node metastasis in patients with a core biopsy diagnosis of ductal carcinoma in situ: an analysis of 255 cases. Breast J 2011;17(3):223–229.
10. Brennan ME, Turner RM, Ciatto S, et al. Ductal carcinoma in situ at core-needle biopsy: meta-analysis of underestimation and predictors of invasive breast cancer. Radiology 2011;260(1):119–128.
11. Byng D, Retèl VP, Schaapveld M, Wesseling J, van Harten WH; Grand Challenge PRECISION consortium. Treating (low-risk) DCIS patients: what can we learn from real-world cancer registry evidence? Breast Cancer Res Treat 2021;187(1):187–196.
12. van Seijen M, Lips EH, Thompson AM, et al. Ductal carcinoma in situ: to treat or not to treat, that is the question. Br J Cancer 2019;121(4):285–292.
13. Grimm LJ, Neely B, Hou R, et al. Mixed-methods study to predict upstaging of DCIS to invasive disease on mammography. AJR Am J Roentgenol 2021;216(4):903–911.
14. Hou R, Mazurowski MA, Grimm LJ, et al. Prediction of upstaged ductal carcinoma in situ using forced labeling and domain adaptation. IEEE Trans Biomed Eng 2020;67(6):1565–1572.
15. Shi B, Grimm LJ, Mazurowski MA, et al. Can occult invasive disease in ductal carcinoma in situ be predicted using computer-extracted mammographic features? Acad Radiol 2017;24(9):1139–1147.
16. Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc Series B Stat Methodol 2010;72(4):417–473.
17. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html.
18. Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK. Learnability and the Vapnik-Chervonenkis dimension. J ACM 1989;36(4):929–965.
19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
20. Bria A, Karssemeijer N, Tortorella F. Learning from unbalanced data: a cascade-based approach for detecting clustered microcalcifications. Med Image Anal 2014;18(2):241–252.
21. Everett MG, Borgatti SP. The centrality of groups and classes. J Math Sociol 1999;23(3):181–201.
22. Hu MK. Visual pattern recognition by moment invariants. IRE Trans Inf Theory 1962;8(2):179–187.
23. Greenwood HI, Wilmes LJ, Kelil T, Joe BN. Role of breast MRI in the evaluation and detection of DCIS: opportunities and challenges. J Magn Reson Imaging 2020;52(3):697–709.
24. Harowicz MR, Saha A, Grimm LJ, et al. Can algorithmically assessed MRI features predict which patients with a preoperative diagnosis of ductal carcinoma in situ are upstaged to invasive breast cancer? J Magn Reson Imaging 2017;46(5):1332–1340.
25. Klimov S, Miligy IM, Gertych A, et al. A whole slide image-based machine learning approach to predict ductal carcinoma in situ (DCIS) recurrence risk. Breast Cancer Res 2019;21(1):83.
26. Kurniawan ED, Rose A, Mou A, et al. Risk factors for invasive breast cancer when core needle biopsy shows ductal carcinoma in situ. Arch Surg 2010;145(11):1098–1104.
27. Park HS, Park S, Cho J, Park JM, Kim SI, Park BW. Risk predictors of underestimation and the need for sentinel node biopsy in patients diagnosed with ductal carcinoma in situ by preoperative needle biopsy. J Surg Oncol 2013;107(4):388–392.
28. Weinmann S, Leo MC, Francisco M, et al. Validation of a ductal carcinoma *in situ* biomarker profile for risk of recurrence after breast-conserving surgery with and without radiotherapy. Clin Cancer Res 2020;26(15):4054–4063.
29. Rakovitch E, Nofech-Mozes S, Hanna W, et al. Multigene expression assay and benefit of radiotherapy after breast conservation in ductal carcinoma in situ. J Natl Cancer Inst 2017;109(4):djw256.
30. Shi B, Grimm LJ, Mazurowski MA, et al. Prediction of occult invasive disease in ductal carcinoma in situ using deep learning features. J Am Coll Radiol 2018;15(3 Pt B):527–534.