

A cross-national study using mammography to predict occult invasive disease in ductal carcinoma in situ

Manuscript Type: Original Research

Summary Statement: Mammographic radiomic features can predict occult invasive disease in core needle biopsy-proven ductal carcinoma in situ, as validated using data from three countries.

Key Results:

- 1,498 women diagnosed with DCIS on core needle biopsy from three countries were included to evaluate models to predict DCIS upstaging.
- A radiomics model trained with mammograms from the United States was comprehensively validated on mammograms from the United Kingdom and the Netherlands.
- The results of this cross-national assessment indicate a generalizability of a predictive radiomics model for DCIS upstaging, with test AUC of 0.604 for the United Kingdom dataset, and 0.682 for the Netherlands dataset.

Abbreviations: DCIS = ductal carcinoma in situ, ER = estrogen receptor, PR = progesterone receptor, ROC = receiver operating characteristic, AUC = area under the curve, PPV = positive predictive value, NPV = negative predictive value, US = United States, UK = United Kingdom, NL = the Netherlands.

ABSTRACT

Background: Patients diagnosed with ductal carcinoma in situ (DCIS) may harbor undetected invasive breast cancer. Radiomic features describing mammographic calcifications can predict occult invasive disease among women diagnosed with DCIS on core needle biopsy, which could change treatment recommendations. Before clinical adoption, the generalizability of these radiomics features must be tested in external validations.

Purpose: To evaluate the generalizability of machine learning models based on radiomic mammography features to predict occult invasive cancer among women diagnosed with DCIS on core needle biopsy from three countries.

Materials and Methods: In this retrospective, cross-national study, digital mammography images from women diagnosed with DCIS on breast core needle biopsy were collected at three sites between January 1, 2000 and December 31, 2021. The study included only asymptomatic women with calcifications but without associated mass, architectural distortion or asymmetry. Logistic regression with regularization was used, and differences across three datasets by upstaging rate, age, lesion size, estrogen and progesterone receptor were assessed by Kruskal-Wallis or Chi-square test.

Results: The study consisted of 1,498 women (age range 31-89, mean 59 ± 9 years), including 696 women from the United States (US), 618 women from the United Kingdom (UK), and 184 women from the Netherlands (NL), with upstaging rates of 16.1%, 16.7%, and 14.1%, respectively. Internal cross-validation AUCs were 0.675 (95%CI: 0.671, 0.679) for the US dataset, 0.603 (95%CI: 0.567, 0.722) for the UK dataset, and 0.701 (95%CI: 0.697, 0.706) for the NL dataset. For external validation, the US-model yielded testing AUC of 0.604 (95% CI: 0.560, 0.648) for the UK dataset, and 0.682 (95%CI: 0.607, 0.757) for the NL dataset.

Conclusion: Radiomic machine learning models demonstrate the potential to predict occult invasive cancer in women with DCIS across diverse settings.

1
2 **INTRODUCTION**
3

4
5 Ductal carcinoma in situ (DCIS) is a stage 0 form of breast cancer which accounts for 15-20% of all new
6 diagnoses (1). Although many patients with DCIS may never progress to invasive ductal carcinoma
7 (IDC) (2-4), standard treatment is surgical excision, because up to 26% of patients with a core needle
8 biopsy diagnosis of DCIS are upstaged to IDC at surgery (5-9) due to undersampling at initial diagnosis.
9 While evaluating all suspicious calcifications for invasive disease prior to biopsy might seem valuable,
10 the next step remains biopsy. This study focuses instead on patients with biopsy-confirmed DCIS, where
11 the critical challenge is to identify those with occult invasion to inform the next decisions in patient
12 management. Alternative management strategies such as active monitoring, which avoids surgery in
13 favor of regular follow-up imaging, are being explored to reduce overtreatment in women diagnosed
14 with low-risk DCIS (10, 11). Since these patients avoid surgery, delayed diagnosis of any occult invasive
15 cancer could lead to progression without appropriate treatment. Therefore, improving the pre-surgical
16 diagnosis of DCIS and occult invasive cancer has important clinical implications that can assist
17 providers and patients in choosing optimal management strategies.
18
19
20
21
22
23
24
25
26
27
28
29
30
31

32
33 Multiple studies have investigated the clinical, pathological, and radiologist-assessed imaging factors
34 associated with DCIS diagnosed at core needle biopsy that is subsequently upstaged to invasive cancer
35 at surgery (9, 12-14). Imaging features include a mass, architectural distortion, asymmetry, and larger
36 lesion sizes, in addition to palpable lesions (13). DCIS with higher nuclear grade is also more likely to
37 be upstaged at excision. Computer-extracted radiomics features have the potential to improve the
38 predictive performance over radiologist-assessed features. In a previous study, digitally extracted
39 mammographic radiomic features derived from a cohort of 700 women demonstrated promise to predict
40 DCIS upstaging (14). The current study seeks to validate the robustness and accuracy of the model in
41 a multi-national setting with more diverse patient populations, clinical settings, and imaging platforms.
42
43
44
45
46
47
48
49
50
51

52
53 We conducted a retrospective, cross-national cohort study to determine the predictive potential of
54 machine learning models using radiomic mammography features to distinguish pure DCIS from DCIS
55
56
57
58
59

with occult invasive cancer. Specifically, the purpose of this study was to validate the robustness of previous radiomics models to predict DCIS upstaging with data from two other countries.

MATERIALS AND METHODS

Study Sample

This retrospective study included three independent datasets, located in the United States (US), the United Kingdom (UK), and the Netherlands (NL). This study was compliant with the Health Insurance Portability and Accountability Act and the Institutional Review Boards from all three institutions approved the study with waivers of informed consent.

Data collection periods varied among the datasets. Cases were collected in the US between September 2008 and April 2017, in the UK between January 2010 and December 2021, and in the NL between January 2008 and May 2020. When available, estrogen receptor (ER), progesterone receptor (PR), and nuclear grade data were recorded from the initial core needle biopsy reports. Inclusion criteria involved patients with mammographic calcifications non-operatively diagnosed as DCIS without invasive cancer, without concomitant contralateral breast cancer or prior history of breast cancer. Exclusion criteria were DCIS presenting as a mass, asymmetry, architectural distortion, or symptomatic patients. Surgical excision pathology reports were reviewed to identify the presence of invasive cancer at surgical excision, i.e., upstaging.

Mammography vendors and imaging protocols varied among the datasets. For the US, all women underwent full-field digital mammography, but the dataset consists of only the magnification views acquired by either Senographe Essential (GE Healthcare) or Selenia Dimensions (Hologic Inc.) systems, with either a 1.5 or 1.8 magnification factor. The UK dataset contained only full-field mammography acquired using Selenia Dimensions (Hologic). Differences in imaging technique were based on each site's standard of care protocols for evaluating suspicious calcifications.

1
2 Biopsy techniques also varied across the three datasets. The US patients underwent 9-gauge vacuum-
3
4 assisted core needle biopsy at a single health system. In the UK, patients from three centers followed
5
6 guidelines recommending 12-pass vacuum-assisted biopsy, though detailed procedural data are
7
8 unavailable. In the Netherlands, standard practice was a 9- or 10-gauge vacuum-assisted biopsy with
9
10 about 6 cores. Occasionally, UK and NL patients underwent 14-gauge biopsy if calcifications were
11
12 clearly visible on ultrasound. All samples in all three countries underwent specimen radiography.
13
14

15
16 A previous study utilized 700 patients from the US dataset to construct a radiomic features' model and
17
18 exclusively evaluated its performance by splitting the data into training and test sets (14). Of these, 140
19
20 patients from the training set were also included in another publication (15), which investigated the
21
22 supplementary roles of atypical ductal hyperplasia and invasive ductal carcinoma in improving
23
24 prediction performance.
25
26

27 ***Radiologic Annotation and Feature Extraction***
28
29

30
31 Each breast lesion was annotated by radiologists with specific expertise in breast imaging. For
32
33 consistency, calcifications within the annotated regions were automatically segmented using a U-Net
34
35 convolutional neural network, which had been trained using images from the US dataset, and clustered
36
37 using a previously reported computer vision algorithm (16).
38
39

40
41 To ensure consistency with previous studies (15, 17), 109 radiomic features were selected to
42
43 characterize both individual calcifications and calcification clusters. These features were chosen for
44
45 their relevance in identifying suspicious calcifications and DCIS upstaging. The selection process
46
47 emphasized shape, morphology, and texture features. To ensure uniformity across vendors and
48
49 institutions, all features underwent pre-processing standardization with zero mean and unit variance
50
51 separately for each vendor (GE or Hologic) and within each country.
52

53 ***Model Building and Testing***
54
55
56
57
58
59
60

Initially, previously built models from the US dataset were validated. Images from the US dataset served as the training set, and a repeated cross-validation approach was undertaken to enhance the training robustness and reliability. Specifically, this iterative process included random shuffling of the training set to mitigate biases stemming from sample ordering across the training folds. Within each iteration, a nested cross-validation structure was implemented, where the outer loop facilitated resampling, while the inner loop conducted hyperparameter tuning and stabilized feature selection (18, 19). After features were selected, repeated cross-validation was implemented again to find the best hyperparameter associated with the selected features. The GridsearchCV function from the Python module, 'scikit-learn 0.20', with default parameters was utilized, and logistic regression with l_2 -regularization (20) was applied as the model classifier. In this classifier, pure DCIS without subsequent upstaging was labeled as negative, whereas DCIS that was upstaged to invasive cancer at the time of surgery was labeled as positive.

A final classifier with predetermined hyperparameters and features was applied, selected based on its highest frequency during repeated cross-validation. For the primary study, the test set consisted of images from the UK and the NL, and the classifier from the US was applied to assess its respective external validation performances. As secondary external validations, we repeated the process by training on the UK dataset and subsequently testing on the other two countries. This was then repeated again by training on the NL dataset and testing on the other two countries. In order to establish a single cutoff, a single model was created using the final classifier from the US. This was applied to the same US dataset to generate two hypothetical operating points, corresponding to 90% sensitivity for active monitoring and 90% specificity for surgical planning. These fixed operating points were then applied to the other two test sets' results to evaluate their performance.

Statistical Analysis

Performance was evaluated by the receiver operating characteristic (ROC) area under the curve (AUC) with 95% confidence intervals (21). Statistics of patient characteristics and biomarkers were analyzed

1 with Chi-square tests for categorical variables and Kruskal-Wallis tests for continuous variables. All
2
3
4 statistical tests adhered to a two-tailed approach with a significance level of 0.05. The implementation
5
6 of feature extraction and machine learning models was executed in Python, while R was employed for
7
8 statistical analyses. The entire codebase is publicly accessible on GitHub
9
10 (<https://github.com/dcis101/multi-national-dcis-study>).
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

RESULTS

Study sample characteristics

There were 1,498 women who met eligibility criteria: 696 women from the US, 618 women from the UK (22), and 184 women from the NL. A flowchart of patients from each site included for analysis is shown in Figure 1. Demographic and pathologic characteristics for each study site are shown in Table 1. Upstaging rates did not differ significantly between the three sites (range: 14.1%-16.7%, $p>0.05$ for all pair-wise comparisons, Chi-square test). However, we noted some differences between datasets: patient age at diagnosis was significantly younger in the NL dataset compared to the other two (56y vs. 59y and 59y; $p<0.05$, Kruskal-Wallis test). Based on the radiomics major axis length (endpoints of the longest line drawn through the detected lesion), lesion size of the UK dataset was significantly smaller than that of the US or NL (20.9 mm vs. 27.3 mm and 26.8 mm, respectively; $p<0.05$, Kruskal-Wallis test). Lesion size differences across three countries are shown as violin plots (Figure 2).

Training by Cross-Validation and Feature Selection

Among the 109 radiomic features analyzed, the top 10% or 11 selected features were identified in the US dataset. These features were the standard deviation, maximum, minimum, and mean of the individual calcifications normalized degree (23); standard deviation of the calcifications mean background intensity; standard deviation and maximum of calcifications distance to the cluster centroid; number of calcifications in the lesion; standard deviation of calcifications mean intensity; maximum of calcifications minor axis; and maximum of calcifications 3rd Hu's moments variant (24). With these 11 features, the US dataset achieved an internally cross-validated AUC of 0.675 (95% CI: 0.671, 0.679).

Cross-National Validation

For the primary external validation, the model trained on US images was tested on the other two countries' images, yielding AUCs of 0.604 (95% CI: 0.560, 0.648) for the UK dataset and 0.682 (95%CI:

1
2 0.607, 0.757) for the NL dataset. Figure 3 depicts ROC curves when the model trained on US images
3
4 tested on the UK and the NL datasets.
5

6
7 To show combined results of the primary and secondary external validations, a confusion matrix was
8
9 generated by taking each row's images as the training set, and each column's images as the test set,
10
11 and repeating the above process. The training of each new model on a new dataset also involved the
12
13 aforementioned procedure of repeated cross-validation with hyperparameter tuning and selection of a
14
15 new set of top 10% of features. Each train-test-performance is shown in Table 2, and the most frequently
16
17 selected features when taking different countries' images as training sets are shown in Figure 4.
18
19

20
21
22 As shown by the middle column in Table 2, performances were consistently lower when all 3 models
23
24 were tested against the UK dataset. In fact, the model trained on UK data performed better when tested
25
26 on data from the US and the NL. Independent of the modeling process, this trend suggests that the UK
27
28 dataset is inherently more challenging than the other two. This also corresponds to earlier observation
29
30 that the UK data showed different distributions in lesion sizes, i.e., the positive upstaged lesions were
31
32 smaller and thus less separable from the similarly sized but negative pure DCIS lesions. Figure 5 shows
33
34 correctly classified examples from both classes: pure DCIS and upstaged DCIS from three different
35
36 datasets.
37

38
39 **Performance Analysis**
40

41
42
43 To compare differences in the performance of clinical management strategies across countries, two
44
45 cutoffs depicting hypothetical operating points were selected based on the US dataset's internal self-
46
47 test model. Note that the internal self-test model for the US dataset was not intended to indicate the
48
49 model's true performance on US data. Instead, it was used solely to set cutoff points, allowing for a
50
51 standardized comparison of performance across external validation datasets and enabling consistent
52
53 evaluation of clinical management strategies across countries. The active monitoring cutoff (upper right
54
55 of ROC curve) was chosen to yield 90% sensitivity and high negative predictive value (NPV).
56
57

Conversely, the surgical planning cutoff (lower left of ROC curve) yielded 90% specificity and high positive predictive value (PPV). The values for sensitivity, specificity, NPV/PPV, and odds ratio are shown in Table 3. Viewing the model through the lens of active monitoring that requires high NPV, the US-trained model produced NPV of 92% and odds ratio of 2.3 on the UK dataset and NPV of 89% and odds ratio of 1.9 on the NL dataset. Conversely, when viewed through the lens of surgical planning to determine if lymph node staging might be indicated, our radiomics model produced PPV of 30% and odds ratio of 2.4 for the UK dataset, and PPV of 30% and odds ratio of 3.2 for the NL dataset. For both active monitoring and surgical planning, note the tight clustering of the three operating points corresponding to each dataset, indicating strong concordance in the cross-country external validations.

DISCUSSION

Classification models for tumor upstaging have the potential to inform clinical management of patients with DCIS and mitigate some of the effects of undersampling by core needle biopsy. It is exceptionally challenging to use mammography images to distinguish patients who share the same biopsy diagnosis of DCIS without invasion. Single-institution studies have shown promise, but this study represents the next step in the adoption pathway by conducting a multi-national external validation study. Although the overall AUC appears modest, the model achieves clinically meaningful performance at key operating points, yielding consistently high NPV for active monitoring and robust PPV for surgical planning, with odds ratios ranging from 1.9 to 3.2 across the two clinical scenarios and two external validation datasets. To put those performances in perspective, the Oncotype DX® DCIS Score is a genomic test that predicts recurrence with hazard ratios ranging from 1.8 to 2.4 for different categories of recurrences (25), and the DCISionRT Decision Score combines immunohistochemistry and clinical features to predict recurrence with hazard ratios ranging from 3.1 to 4.2 (26).

Previous studies in predicting DCIS upstaging have several key limitations (9, 13-15, 27, 28). First, many studies have used relatively small sample size, often with less than 100 patients. Greater sample

1 sizes are typically needed for more advanced radiomics analysis which leverages large numbers to
2 identify relevant findings. Our study of nearly 1,500 patients is the largest published series of DCIS
3 patients to date. The substantial size of our dataset enhances the robustness of our findings and allows
4 for more comprehensive analyses. Second, most published studies have focused on single-institution
5 datasets for testing and validation, which is much more practicable but performance often drops when
6 models are applied to other locations (29, 30). Patient demographics, pathologic distributions of disease,
7 and imaging techniques often differ across institutions and between countries. Furthermore, there are
8 also notable differences in practice patterns. The biopsy recommendations in the US and the UK are
9 based on diagnostic magnification views, while in the NL only screening FFDM images are utilized.
10 Screening in the UK and the NL use double reading, but single reading is the standard in the US. In
11 contrast, our study used a multi-site, cross-national approach which supports the generalizability of our
12 approach. Third, in addition to reductions in performance, the selection of features by radiomics models
13 often differs across sites. In contrast, the observed consistency in both selected features as well as
14 performance metrics across different countries underscores the reliability of our approach. This internal
15 coherence among diverse datasets enhances the credibility and applicability of radiomic features DCIS,
16 supporting the robustness of our methodology across varied healthcare settings. Moreover, the
17 observed variations in imaging practices, biopsy techniques, and patient populations reflect real-world
18 clinical diversity, which our study aimed to capture. Despite these differences, the model demonstrated
19 consistent performance across datasets, supporting its potential generalizability, which is an essential
20 step toward broader applicability.

21 Typically, the performance of radiomics analyses is distilled down to the AUC. While this metric
22 assesses the overall performance and facilitates comparisons with other studies, it lacks real world
23 relevance and limits the ability to translate this work to clinical practice. Our study investigates two
24 different clinically relevant scenarios: high sensitivity and high specificity. The high-specificity cutoff
25 optimizes PPV to identify patients at high likelihood for occult invasive cancer. Not only would these
26 patients be poor active monitoring candidates, but surgeons may also wish to change their surgical plan

to sample axillary lymph nodes given the higher likelihood of upstaging. Conversely, the high-sensitivity cutoff maximizes NPV to identify patients who would be good candidates for active monitoring. Our results demonstrated a sensitivity of 85-90% and NPV of 89-93% across datasets. The success or failure of active monitoring will depend in large part on correct patient selection, so it is paramount to be able to tune the model to perform under specific clinical circumstances. These radiomics models may recommend different women compared to the clinical criteria for current active monitoring trials; such differences remain to be reconciled.

This study had several limitations. Firstly, variations in upstaging rates among three countries represent a potential confounding factor that should be analyzed. These differences may be caused by factors such as each country's patient population, diagnostic practices, or other unaccounted for variables. Secondly, the presence of mammography vendor-related disparities introduces a layer of complexity, as different imaging systems may exhibit nuanced variations in performance characteristics. Thus, in our study, we intentionally excluded certain images sourced from vendors with limited representation, as their small sample size could not adequately reflect the diversity of the overall population. Thirdly, the NL and UK datasets did not include certain clinicopathologic features, which limits validation of other previously reported models. Such dataset heterogeneity is commonly encountered during external validation. Our present focus is a mammography-only approach, which offers a non-invasive option crucial for women undergoing active monitoring for multiple years. Future studies may incorporate additional features, and evaluate alternative modeling approaches including deep learning to further improve predictive performance. In addition, more balanced datasets will be collected in the NL side to further validate performance across diverse populations. Lastly, our study observed that the UK dataset exhibited the lowest performance in both training and testing, indicating inherent challenges specific to that dataset. A possible explanation lies in the different distributions in lesion sizes in that UK dataset, which reduced separability between the classes of pure DCIS versus upstaged. However, in spite of these dataset differences, all models showed promising generalizability. Even the UK-trained model performed better when externally validated on the other two datasets. These disparities suggest

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

potential nuances in lesion characteristics or imaging protocols, which underscore the need for cautious interpretation and highlight areas for further investigation and refinement in the future.

In conclusion, our comprehensive investigation into the predictive potential of radiomic features in DCIS provides a deeper understanding of the intricate features associated with DCIS biology and has potential for clinical utility. The cross-national analysis, encompassing diverse clinical contexts, enhances the generalizability of our findings and the robustness of our methodology across different healthcare settings. These collective findings hold promise for improving pre-surgical diagnosis and management strategies for DCIS. As we navigate the complex landscape of breast cancer diagnostics, our study offers a foundation for future research endeavors and clinical applications in the realm of radiomic feature analysis.

References

1. American Cancer Society. Breast Cancer Facts & Figures 2019-2020. Atlanta: American Cancer Society, Inc. 2019.
2. 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. doi: 10.1002/cncr.21069
3. Hieken TJ, Krishnamurthy V, Farolan M, Velasco JM. Long-term outcome of DCIS patients: p53 as a biomarker of ipsilateral recurrence. *Journal of Clinical Oncology* 2011;29(27):39-39. doi: DOI 10.1200/jco.2011.29.27_suppl.39
4. Duffy SW, Dibden A, Michalopoulos D, Offman J, Parmar D, Jenkins J, Collins B, Robson T, Scorfield S, Green K, Hall C, Liao X-H, Ryan M, Johnson F, Stevens G, Kearins O, Sellars S, Patnick J. Screen detection of ductal carcinoma in situ and subsequent incidence of invasive interval breast cancers: a retrospective population-based study. *The Lancet Oncology* 2016;17(1):109-114. doi: 10.1016/S1470-2045(15)00446-5
5. Sim YT, Litherland J, Lindsay E, Hendry P, Brauer K, Dobson H, Cordiner C, Gagliardi T, Smart L. 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. doi: 10.1016/j.crad.2014.12.019
6. 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. doi: 10.1016/S1015-9584(10)60013-9
7. Mannu GS, Wang Z, Broggio J, Charman J, Cheung S, Kearins O, Dodwell D, Darby SC. 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. doi: 10.1136/bmj.m1570
8. 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. doi: 10.1111/j.1524-4741.2011.01069.x
9. Schulz S, Sinn P, Golatta M, Rauch G, Junkermann H, Schuetz F, Sohn C, Heil J. Prediction of underestimated invasiveness in patients with ductal carcinoma in situ of the breast on percutaneous

1
2 biopsy as rationale for recommending concurrent sentinel lymph node biopsy. Breast 2013;22(4):537-
3 542. doi: 10.1016/j.breast.2012.11.002
4
5
6 10. Glencer AC, Miller PN, Greenwood H, Maldonado Rodas CK, Freimanis R, Basu A, Mukhtar RA,
7 Brabham C, Kim P, Hwang ES, Rosenbluth JM, Hirst GL, Campbell MJ, Borowsky AD, Esserman LJ.
8 Identifying Good Candidates for Active Surveillance of Ductal Carcinoma In Situ: Insights from a
9 Large Neoadjuvant Endocrine Therapy Cohort. Cancer Res Commun 2022;2(12):1579-1589. doi:
10 10.1158/2767-9764.Crc-22-0263
11
12
13
14
15 11. Van Bockstal MR, Agahozo MC, Koppert LB, van Deurzen CHM. A retrospective alternative for
16 active surveillance trials for ductal carcinoma in situ of the breast. International Journal of Cancer
17 2020;146(5):1189-1197. doi: <https://doi.org/10.1002/ijc.32362>
18
19
20
21 12. Meurs CJC, van Rosmalen J, Menke-Pluijmers MBE, Ter Braak BPM, de Munck L, Siesling S,
22 Westenend PJ. A prediction model for underestimation of invasive breast cancer after a biopsy
23 diagnosis of ductal carcinoma in situ: based on 2892 biopsies and 589 invasive cancers. Br J Cancer
24 2018;119(9):1155-1162. doi: 10.1038/s41416-018-0276-6
25
26
27
28 13. Brennan ME, Turner RM, Ciatto S, Marinovich ML, French JR, Macaskill P, Houssami N. Ductal
29 Carcinoma in Situ at Core-Needle Biopsy: Meta-Analysis of Underestimation and Predictors of
30 Invasive Breast Cancer. Radiology 2011;260(1):119-128. doi: 10.1148/radiol.11102368
31
32
33
34 14. Hou R, Grimm LJ, Mazurowski MA, Marks JR, King LM, Maley CC, Lynch T, van Oirsouw M,
35 Rogers K, Stone N, Wallis M, Teuwen J, Wesseling J, Hwang ES, Lo JY. Prediction of Upstaging in
36 Ductal Carcinoma in Situ Based on Mammographic Radiomic Features. Radiology 2022;303(1):54-62.
37 doi: 10.1148/radiol.210407
38
39
40
41 15. Hou R, Mazurowski MA, Grimm LJ, Marks JR, King LM, Maley CC, Hwang ES, Lo JY. Prediction
42 of Upstaged Ductal Carcinoma In Situ Using Forced Labeling and Domain Adaptation. IEEE Trans
43 Biomed Eng 2020;67(6):1565-1572. doi: 10.1109/TBME.2019.2940195
44
45
46
47 16. Hou R, Peng Y, Grimm LJ, Ren Y, Mazurowski MA, Marks JR, King LM, Maley CC, Hwang ES, Lo
48 JY. Anomaly Detection of Calcifications in Mammography Based on 11,000 Negative Cases. IEEE
49 Trans Biomed Eng 2022;69(5):1639-1650. doi: 10.1109/tbme.2021.3126281
50
51
52
53 17. Bria A, Karssemeijer N, Tortorella F. Learning from unbalanced data: a cascade-based approach
54 for detecting clustered microcalcifications. Med Image Anal 2014;18(2):241-252. doi:
55 10.1016/j.media.2013.10.014
56
57
58
59
60

18. Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2010;72(4):417-473.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011;12:2825-2830.
20. Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK. Learnability and the Vapnik-Chervonenkis Dimension. *Journal of the Acm* 1989;36(4):929-965. doi: Doi 10.1145/76359.76371
21. 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.
22. Halling-Brown MD, Warren LM, Ward D, Lewis E, Mackenzie A, Wallis MG, Wilkinson LS, Given-Wilson RM, McAvinchey R, Young KC. OPTIMAM Mammography Image Database: A Large-Scale Resource of Mammography Images and Clinical Data. *Radiol Artif Intell* 2021;3(1):e200103. doi: 10.1148/ryai.2020200103
23. Everett MG, Borgatti SP. The centrality of groups and classes. *Journal of Mathematical Sociology* 1999;23(3):181-201. doi: Doi 10.1080/0022250x.1999.9990219
24. Hu M. Visual-Pattern Recognition by Moment Invariants. *Ire Transactions on Information Theory* 1962;8(2):179-&. doi: DOI 10.1109/tit.1962.1057692
25. Rakovitch E, Nofech-Mozes S, Hanna W, Baehner FL, Saskin R, Butler SM, Tuck A, Sengupta S, Elavathil L, Jani PA, Bonin M, Chang MC, Robertson SJ, Slodkowska E, Fong C, Anderson JM, Jamshidian F, Miller DP, Cherbavaz DB, Shak S, Paszat L. A population-based validation study of the DCIS Score predicting recurrence risk in individuals treated by breast-conserving surgery alone. *Breast Cancer Res Treat* 2015;152(2):389-398. doi: 10.1007/s10549-015-3464-6
26. Bremer T, Whitworth PW, Patel R, Savala J, Barry T, Lyle S, Leesman G, Linke SP, Jirström K, Zhou W, Amini R-M, Wärnberg F. A Biological Signature for Breast Ductal Carcinoma In Situ to Predict Radiotherapy Benefit and Assess Recurrence Risk. *Clinical Cancer Research* 2018;24(23):5895-5901. doi: 10.1158/1078-0432.Ccr-18-0842
27. Shi B, Grimm LJ, Mazurowski MA, Baker JA, Marks JR, King LM, Maley CC, Hwang ES, Lo JY. Can Occult Invasive Disease in Ductal Carcinoma In Situ Be Predicted Using Computer-extracted Mammographic Features? *Acad Radiol* 2017;24(9):1139-1147. doi: 10.1016/j.acra.2017.03.013

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

28. Zhu Z, Harowicz M, Zhang J, Saha A, Grimm LJ, Hwang ES, Mazurowski MA. Deep learning analysis of breast MRIs for prediction of occult invasive disease in ductal carcinoma in situ. *Comput Biol Med* 2019;115:103498. doi: 10.1016/j.compbimed.2019.103498

29. Wang X, Liang G, Zhang Y, Blanton H, Bessinger Z, Jacobs N. Inconsistent Performance of Deep Learning Models on Mammogram Classification. *Journal of the American College of Radiology* 2020;17(6):796-803. doi: <https://doi.org/10.1016/j.jacr.2020.01.006>

30. Vries CFd, Colosimo SJ, Staff RT, Dymiter JA, Yearsley J, Dinneen D, Boyle M, Harrison DJ, Anderson LA, Lip G, Collaboration obotiR, Black C, Murray AD, Wilde K, Blackwood JD, Butterly C, Zurowski J, Eilbeck J, McSkimming C. Impact of Different Mammography Systems on Artificial Intelligence Performance in Breast Cancer Screening. *Radiology: Artificial Intelligence* 2023;5(3):e220146. doi: 10.1148/ryai.220146

TABLES

Table 1. Patient and DCIS Characteristics by Study Site.

Characteristic	US	UK	NL	p-value
Upstaged				0.713**
No. of upstaged DCIS	112	103	26	
Upstaging rate	16.1%	16.7%	14.1%	
Age at diagnosis (years)				<0.05*
Average	59	59	56	
Range	40-89	41-82	31-77	
Lesion size (mm)				<0.05*
Average	27.3	20.9	26.8	
Range	1.4-184.6	1.7-151.1	2.5-129.2	
Estrogen receptor status				NA
Positive	561 (80.6%)	313 (50.7%)	NA	
Negative	129 (18.5%)	57 (9.2%)	NA	
Unknown	6 (0.9%)	248 (40.1%)	NA	
Progesterone receptor status				NA
Positive	496 (71.3%)	178 (28.8%)	NA	
Negative	193 (27.7%)	62 (10.0%)	NA	
Unknown	7 (1.0%)	378 (61.2%)	NA	
Nuclear grade				NA
Low (Grade I)	26 (3.7%)	18 (2.9%)	42 (22.8%)	
Intermediate (Grade II)	250 (35.9%)	64 (10.4%)	84 (45.7%)	
High (Grade III)	420 (60.4%)	13 (2.1%)	58 (31.5%)	
Unknown	0 (0%)	523 (84.6%)	0	

**Chi-square test

*Kruskal-Wallis test

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 2. Confusion Matrix of AUCs with Different Training and Test sets. The diagonal represents internal cross-validation for each country. Off-diagonal cells represent external validations. 95% confidence intervals are shown in brackets.

		Testing		
		US	UK	NL
Training	US	0.675 [0.671-0.679]	0.604 [0.560-0.648]	0.682 [0.607-0.757]
	UK	0.665 [0.609-0.721]	0.603 [0.567-0.632]	0.644 [0.561-0.715]
	NL	0.632 [0.575-0.689]	0.582 [0.538-0.627]	0.701 [0.697-0.706]

Table 3. Performance of Clinical Management Strategies Across Countries. The active monitoring cutoff (upper right of US ROC curve) was chosen to yield 90% sensitivity and high NPV. Conversely, the surgical planning cutoff (lower left of US ROC curve) yielded 90% specificity and high PPV.

Practice	Statistics	US	UK	NL
Active Monitoring	Sensitivity	90%	86%	85%
	Specificity	27%	24%	21%
	NPV	93%	89%	92%
	Odds Ratio	3.33	1.89	2.32
Surgical Planning	Sensitivity	26%	27%	20%
	Specificity	90%	90%	90%
	PPV	33%	30%	30%
	Odds Ratio	3.16	3.21	2.38

1
2 **FIGURE LEGENDS**
3

4
5 **Figure 1. Flowchart of patients with DCIS included in the study.**
6
7

8
9
10 **Figure 2. DCIS Lesion Size Differences for the Three Study Sites.** In each violin plot, left half
11 (brighter color) corresponds to negative or pure DCIS, while right half (faded color) shows positive or
12 upstaged patients. For all three datasets, pure DCIS (left half) consistently has smaller size distribution
13 than upstaged (right half), but the UK dataset has less separation between its two halves. The area of
14 each half is normalized to emphasize the relative distribution rather than absolute number of patients.
15
16
17
18
19
20
21

22
23 **Figure 3. Primary External Validation Cutoffs of Active Monitoring and Surgical Planning.**
24

25 Model was trained on images from the US and tested on images from the UK and the NL. Note the
26 similar operating points and clinically promising performances for active monitoring (upper right) and
27 surgical planning (lower left).
28
29
30
31

32
33
34 **Figure 4. Stacked Frequency of Top Selected Training Features.** During repeated cross-validation,
35 a y-axis value of 1 corresponded to that feature being selected 100% of the time for each country, up
36 to a total value of 3 across the 3 countries. Persistence of color bands indicate many features were
37 selected consistently regardless of which dataset was used for training. The top 40 of the 109 features
38 are shown.
39
40
41
42
43
44
45
46
47

48 **Figure 5. Examples of correctly predicted lesions among three countries using the model trained**
49 **on images from the US.** Lesion annotations were provided by the study radiologists at each site.
50
51
52
53
54
55
56
57
58
59
60

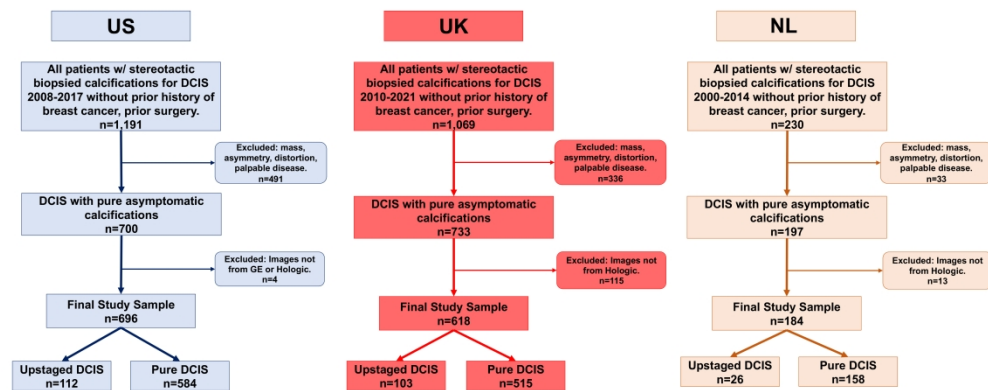


Figure 1. Flowchart of patients with DCIS included in the study.

482x208mm (300 x 300 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

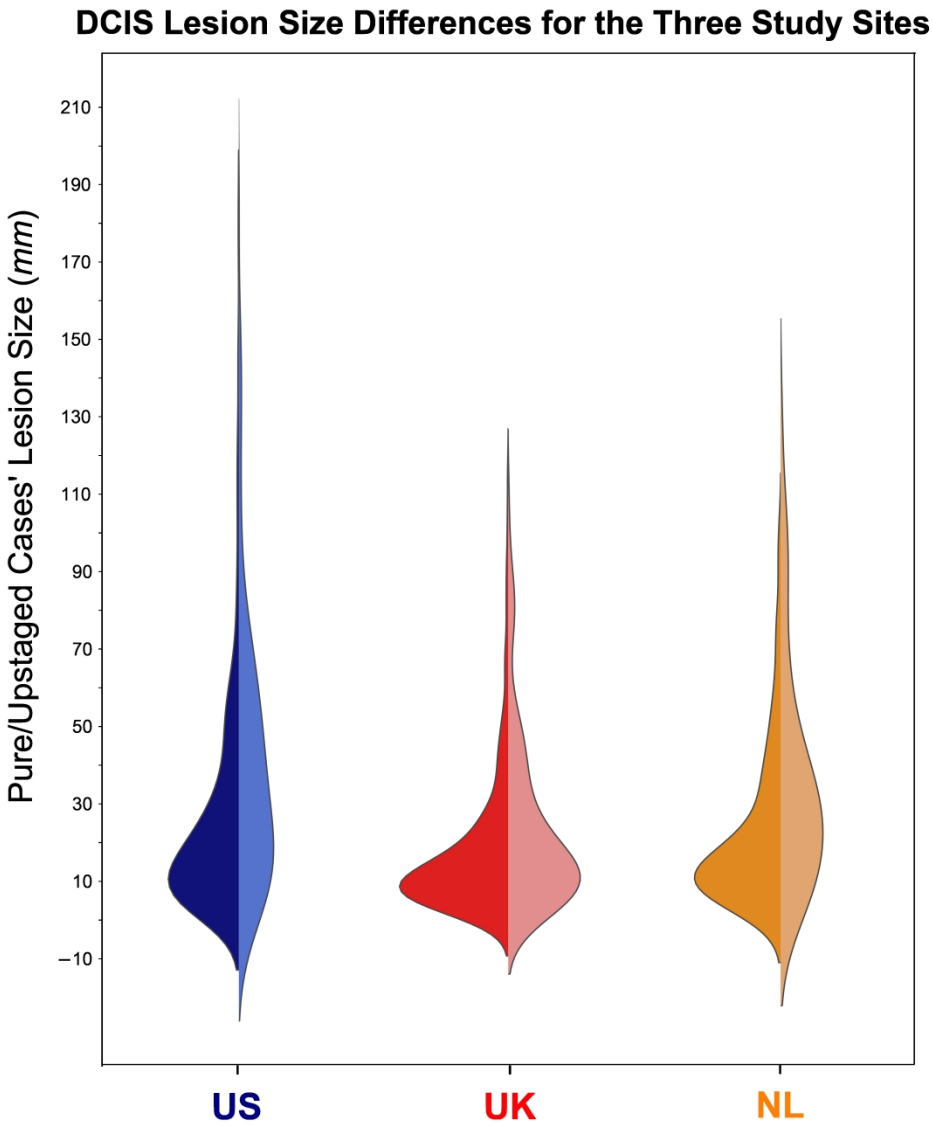


Figure 2. DCIS Lesion Size Differences for the Three Study Sites. In each violin plot, left half (brighter color) corresponds to negative or pure DCIS, while right half (faded color) shows positive or upstaged patients. For all three datasets, pure DCIS (left half) consistently has smaller size distribution than upstaged (right half), but the UK dataset has less separation between its two halves. The area of each half is normalized to emphasize the relative distribution rather than absolute number of patients.

386x443mm (300 x 300 DPI)

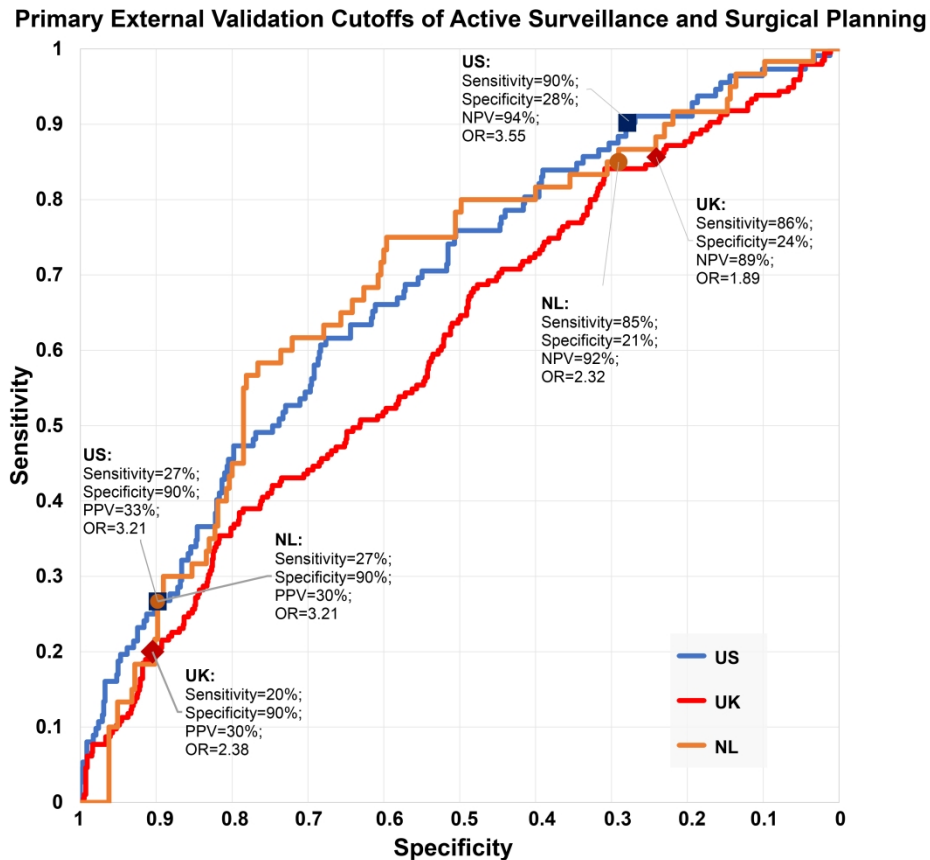


Figure 3. Primary External Validation Cutoffs of Active Monitoring and Surgical Planning. Model was trained on images from the US and tested on images from the UK and the NL. Note the similar operating points and clinically promising performances for active monitoring (upper right) and surgical planning (lower left).

353x304mm (300 x 300 DPI)

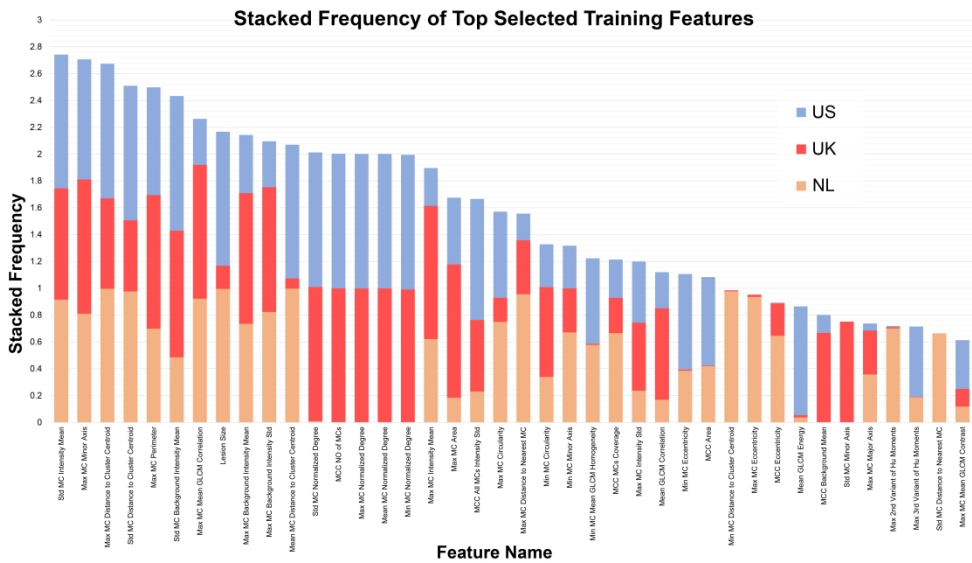


Figure 4. Stacked Frequency of Top Selected Training Features. During repeated cross-validation, a y-axis value of 1 corresponded to that feature being selected 100% of the time for each country, up to a total value of 3 across the 3 countries. Persistence of color bands indicate many features were selected consistently regardless of which dataset was used for training. The top 40 of the 109 features are shown.

482x279mm (300 x 300 DPI)

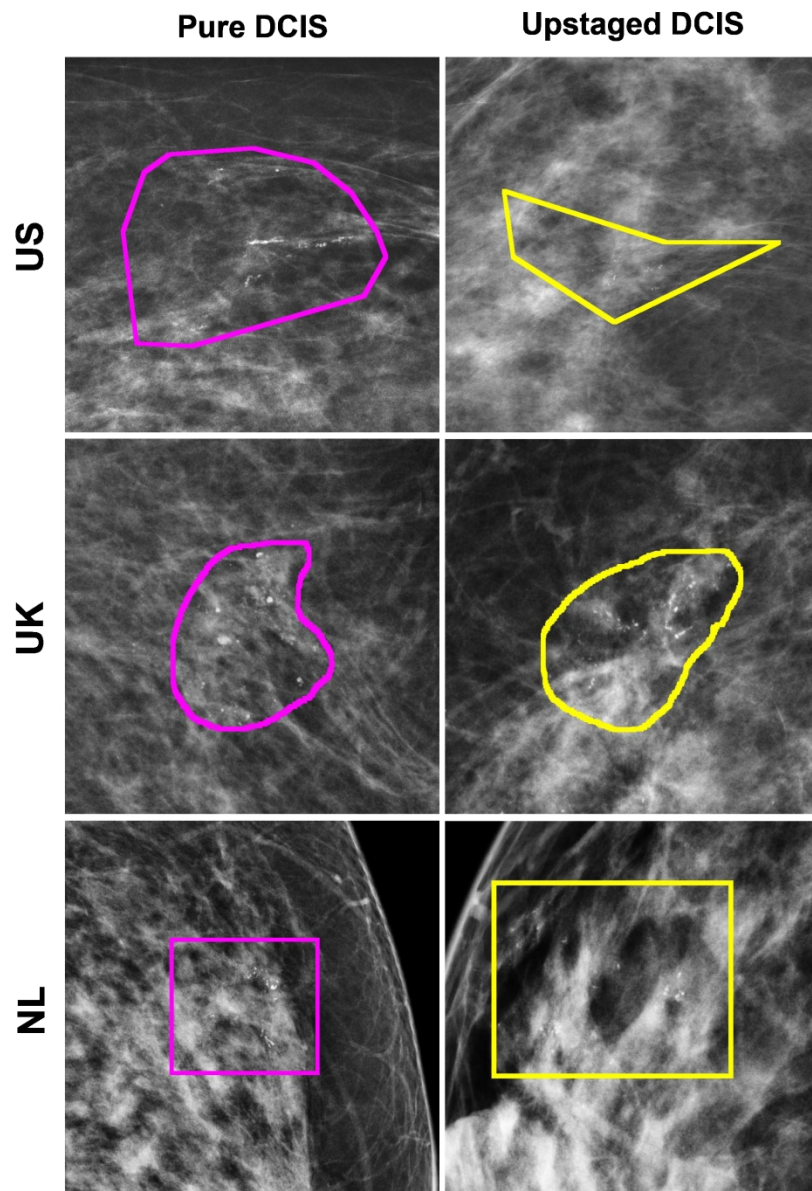


Figure 5. Examples of correctly predicted lesions among three countries using the model trained on images from the US. Lesion annotations were provided by the study radiologists at each site.

187x273mm (300 x 300 DPI)

STARD 2015 Checklist

Manuscript type: Original Research
Manuscript title: A cross-national study using mammography to predict occult invasive disease in ductal carcinoma in situ

Section & Topic	No.	Item	✓ (N/A)
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	✓
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	✓
INTRODUCTION			
	3	Scientific & clinical background, including the intended use and clinical role of the index test	✓
	4	Study objectives and hypotheses	✓
METHODS			
Study design	5	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	✓
Participants	6	Eligibility criteria	✓
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	✓
	8	Where and when potentially eligible participants were identified (setting, location and dates)	✓
	9	Whether participants formed a consecutive, random or convenience series	✓
Test methods	10a	Index test, in sufficient detail to allow replication	✓
	10b	Reference standard, in sufficient detail to allow replication	✓
	11	Rationale for choosing the reference standard (if alternatives exist)	✓
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	✓
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	✓
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	✓
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	✓
Analysis	14	Methods for estimating or comparing measures of diagnostic accuracy	✓
	15	How indeterminate index test or reference standard results were handled	✓
	16	How missing data on the index test and reference standard were handled	N/A
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	✓
	18	Intended sample size and how it was determined	✓
RESULTS			
Participants	19	Flow of participants, using a diagram. Include the figure number (preferably figure 1) or page number	✓
	20	Baseline demographic and clinical characteristics of participants	✓
	21a	Distribution of severity of disease in those with the target condition	N/A
	21b	Distribution of alternative diagnoses in those without the target condition	N/A
	22	Time interval and any clinical interventions between index test and reference standard	N/A
Test results	23	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	✓
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	✓
	25	Any adverse events from performing the index test or the reference standard	N/A

DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalizability	✓
	27	Implications for practice, including the intended use and clinical role of the index test	✓
OTHER INFORMATION			
	28	Registration number and name of registry	N/A
	29	Where the full study protocol can be accessed	N/A
	30	Sources of funding and other support; role of funders	✓

*N/A stands for not applicable and may be a reasonable choice depending on the type of study performed





DSPACE

<https://dspace.org/>

Cross-national radiomics validation using mammography to predict occult invasion in ductal carcinoma in situ

Hou, Rui; Grimm, Lars J.; Marks, Jeffrey R.; King, Lorraine M.; Lynch, Thomas ...

2025-11-01

Radiological Society of North America (RSNA)

<https://doi.org/10.1148/radiol.243739>

Hou R, Grimm LJ, Marks JR, et al., (2025) Cross-national radiomics validation using mammography to predict occult invasion in ductal carcinoma in situ. Radiology, Volume 317, Issue 2, November 2025, Article number e243739