EUROPEAN JOURNAL OF CANCER 48 (2012) 30–36

Available at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.ejconline.com

Current Perspective

Scoring to predict the possibility of upgrades to malignancy in atypical ductal hyperplasia diagnosed by an 11-gauge vacuum-assisted biopsy device: An external validation study

S. Bendifallah [a,b,*], S. Defert [c], N. Chabbert-Buffet [a], N. Maurin [d], J. Chopier [d], M. Antoine [e], C. Bezu [a], D. Touche [f], S. Uzan [a,e], O. Graesslin [c], R. Rouzier [a,b,g]

[a] Department of Obstetrics and Gynecology, Tenon APHP University Hospital, 75020 Paris, France
[b] ER2, Pierre and Marie Curie University, Paris, France
[c] Department of Obstetrics and Gynaecology, Institute Alix de Champagne University Hospital, 51092 Reims, France
[d] Department of Pathology, Tenon APHP University Hospital, Paris, France
[e] Department of Radiology, Tenon APHP University Hospital, Paris, France
[f] Breast Unit, Institute Jean Godinot, Reims, France
[g] INSERM-UMR_S 938, Pierre and Marie Curie University, Paris, France

ARTICLE INFO

Article history:
Available online 17 November 2011

Keywords:
Atypical ductal hyperplasia
Mathematical models
Underestimation rate
Upgrading rate
Scoring system
Vacuum-assisted biopsy

ABSTRACT

Background: Ko's scoring system was developed to predict malignancy upgrades in patients diagnosed with atypical ductal hyperplasia by core needle biopsy. The Ko algorithm was able to identify a subset of patients who were eligible for exclusively clinical follow-up. The current study statistically investigated the patient outcomes to determine whether this scoring system could be translated and used safely in clinical practice.

Methods: We tested the statistical performance of the Ko scoring system against an external independent multicentre population. One hundred and seven cases of atypical ductal hyperplasia diagnosed by an 11-gauge biopsy needle were available for inclusion in this study. The discrimination, calibration and clinical utility of the scoring system were quantified. In addition, we tested the underestimation rate, sensitivity, specificity, and positive and negative predictive values according to the score threshold.

Results: The overall underestimation rate was 19% (20/107). The area under the receiver operating characteristic curve for the logistic regression model was 0.51 (95% confidence interval: 0.47–0.53). The model was not well calibrated. The lowest predicted underestimation rate was 11%. The sensitivity, specificity, positive predictive value, and negative predictive values were 90%, 22%, 20%, and 89%, respectively, according to the most accurate threshold proposed in the original study.

Conclusion: The scoring system was not sufficiently accurate to safely define a subset of patients who would be eligible for follow-up only and no additional treatment. These results demonstrate a lack of reproducibility in an external population. A multidisciplinary approach that correlates clinicopathological and mammographic features should be recommended for the management of these patients.

* *Corresponding author:* Tel.: +33 (0)1 56 01 68 76/68 49; fax: +33 (0)1 56 01 60 62.
E-mail address: sofiane.bendifallah@yahoo.fr (S. Bendifallah).
0959-8049/$ - see front matter
doi:10.1016/j.ejca.2011.08.011

1. Introduction

Population-based mammography screening has resulted in increased detection of suspicious, non-palpable lesions that require further histopathological assessment. Ultrasound-guided needle biopsy (14–16 gauge) or vacuum-assisted (11–14 gauge) breast biopsy (VABB) systems have become widely-used alternatives to open surgical biopsy.[1–3] Atypical ductal hyperplasia (ADH) of the breast, which is discovered in 2–11%[4,5] of cases, is histopathologically defined as either (i) a hyperplastic lesion with some cytological features of low-grade ductal carcinoma in situ (DCIS) that lacks the overall characteristic architectural growth pattern of DCIS, or (ii) a lesion with the classic cytological and architectural features of low-grade DCIS that is confined to ducts and measures less than 2 mm.[6] The difficulty in achieving acceptable levels of concordance between pathology results from image-guided biopsy (IGB) and surgical excision is a major practical concern.[7,8] Due to the risk of underestimating or upgrading the diagnosis (meaning that DCIS or invasive cancer are present), surgical excision is an accepted option for all women diagnosed with ADH. Various strategies have been unsuccessfully developed to improve cancer detection and the risk of underestimation, including revising the definition criteria, changing the device size, and testing clinical, radiological or pathological factors.[9–13] A risk of upgrade of 2% or less has been suggested by the American College of Radiology[14] to be safe for proposing exclusive follow-up breast imaging in certain cases. Based on a population of ADH cases diagnosed by ultrasound-guided core needle biopsy (CNB), in 2007, Ko et al.[15] developed a logistic regression model as an algorithm for scoring the possibility of predicting malignancy upgrade using a combination of five independent factors. The accuracy of the model was tested. The area under (AUC) the receiver operating characteristic (ROC) curve was 0.90 (95% confidence interval, 0.83–0.97) and 0.85 (95% CI, 0.74–0.95) in the study (74 patients) and validation datasets (54 patients), respectively. Because the study indicated a reported sensitivity and negative predictive value of 100% and no cases were upgraded amongst those with a score of 3.5 or less, the authors concluded that this subset of patients should be eligible for non-invasive management.[15] These relevant findings support the hypothesis that the scoring system should be applicable to another population. To our knowledge, no external validation of this tool has been published. Therefore, in the current report, we evaluated the performance and clinical utility of Ko's scoring system[15] in our population of samples from 11-gauge VABB to determine whether this system could be used in clinical practice.

2. Materials and methods

2.1. Data selection

A multicentre search of the medical databases at Tenon APHP University Hospital and the Institute Alix de Champagne University Hospital to identify only ADH cases diagnosed by imaging-guided biopsies (11-gauge vacuum-assisted biopsy device) and followed by surgical excision between January 2003 and December 2010 revealed 229 cases. Amongst these, 13 cases in which the ADH was associated with malignant lesions (i.e. invasive carcinoma or DCIS) upon biopsy and 109 cases in which the absence of one relevant (radiographic, pathological or clinical) criterion prevented the use of the Ko nomogram for scoring were excluded. The details regarding the missing parameters are reported in Fig. 1. Therefore, 107 (46.7%) cases were eligible for the current validation study. Demographical data, imaging, biopsy and open surgical pathology results were collected for each patient (Table 1).

2.2. Mammography and biopsy evaluation

Biopsy procedures were performed for mammographically detected microcalcifications ($n = 100$), architectural distortions and/or suspicious asymmetric densities ($n = 6$) and palpable mass lesions ($n = 1$) by breast intervention radiologists using an 11-gauge vacuum-assisted biopsy device (Mammotome; Ethicon Endo-Surgery, Cincinnati, Ohio). Each pre-biopsy mammogram was independently reviewed to categorise the lesions according to the mammographic BI-RADS category[14] (categories 3–5) and to measure the maximum mammographic lesion diameter. The percutaneous biopsy specimens and surgery slides of the 107 cases were collected from the electronic medical records and specifically re-reviewed for this analysis. The histological slides were interpreted at either Tenon's University Hospital or at the Institute Alix de Champagne University Hospital by experienced pathologists, and were diagnosed according to the diagnostic criteria of the revised 2003 World Health Organization guidelines and classification.[5] Lesions yielding ADH at biopsy and DCIS or carcinoma at surgery were recorded as ADH underestimations.

All excisions were guided by preoperative wire-localisation, and pathological analysis was performed on the retrieved vacuum cavity. The absence of malignancy was considered to be coincidental and therefore predictable by the Ko model. This consideration does not warrant the continuum between ADH and malignancy but is important to test the accuracy of the Ko model.

2.3. Development of the Ko scoring system

The five independent predictors of malignancy included in the scoring system were selected based on P values <0.05 in the multivariate analyses. The multiple of 0.5 nearest to the β coefficient obtained for each significant factor from the multivariate logistic regression model was assigned for each factor to build the algorithm. A score of 2.0 was assigned for a palpable lesion and microcalcification on mammography; 3.5 was assigned for calcifications >1.5 cm, focal ADH ($\leqslant$1 duct and $\leqslant$1 mm), and age >50 years. The scores for each significant factor were then added, resulting in a total score for each patient. The final scores ranged from 0 to 14.5. A score $\leqslant$3.5 designated a subset of patients with ADH lesions that were defined as 'probably benign' and could be safely followed non-operatively rather than surgically excised.[15]

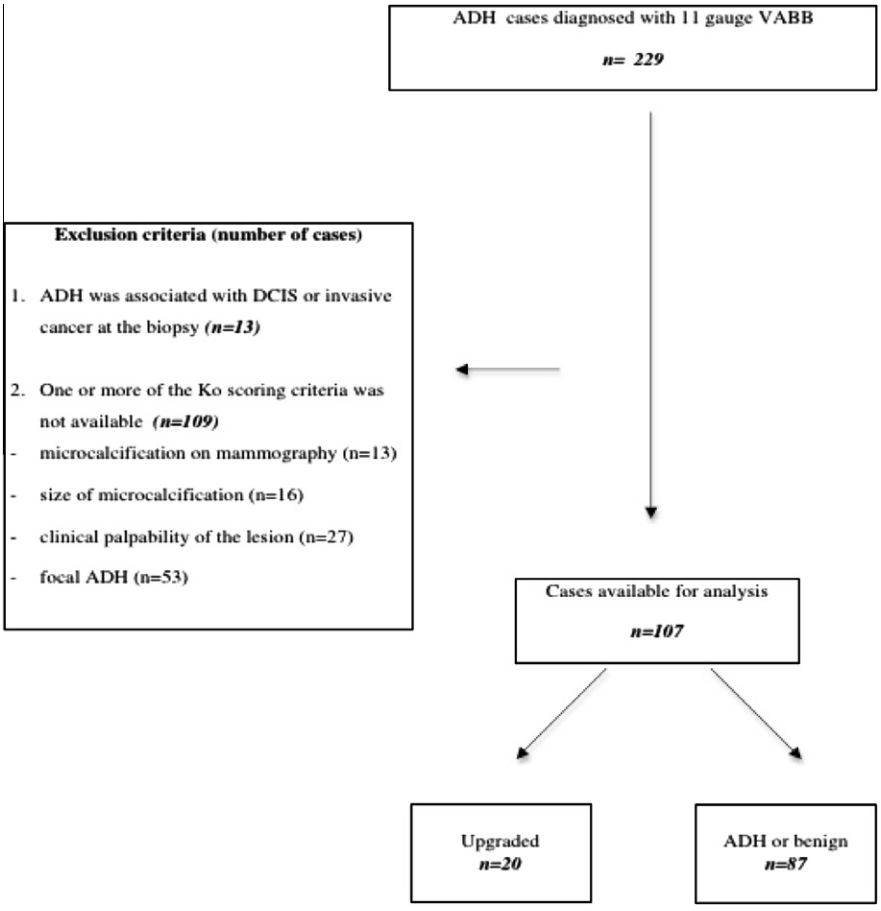

Fig. 1 – Process for data selection from atypical ductal hyperplasia (ADH) cases diagnosed between January 2003 and December 2010.

2.4. *Performance of the logistic regression model (LRM)*

To test the performance of the LRM in our population of 107 patients, we used the following validated statistical tools: discrimination, calibration, and clinical utility.[16]

Discrimination (i.e. whether the relative ranking of individual predictions was in the correct order) was quantified using the AUC, which can range from 0 to 1 (1 indicating perfect concordance, 0.5 indicating no association, and 0 indicating perfect discordance), and its 95% CI.[17] The ROC curves were constructed using the Hanley and McNeil method and showed the relationship between the sensitivity and false-positive rate (1-specificity) of a test across all possible threshold values.[17]

Calibration (i.e. the agreement between the observed outcome frequencies and the predicted probabilities) was analysed from the graphical representation of the relationship between the observed outcome frequencies and the predicted probabilities (calibration curves). A calibration curve can be approximated by a regression line with intercept α and slope β. Well-calibrated models have $\alpha = 0$ and $\beta = 1$. Therefore, a sensible measure of calibration is a likelihood ratio statistic that tests the null hypothesis that $\alpha = 0$ and $\beta = 1$. The statistic has a χ^2 distribution with 2 *df* (unreliability [U] statistic).[18] We also evaluated the average (E average [E_{aver}]) errors between the predictions and observations obtained from the calibration curve.

Clinical utility of the model allows the identification of the largest subgroup of patients with the lowest underestimation rate according to the score value.[16]

2.5. *Performance of the scoring system*

In addition, we tested the underestimation rate, sensitivity, specificity, and positive and negative predictive values according to the different threshold values defined by Ko et al.[15]

The data were analysed with R package version 2.10.1 using the Design, Hmisc and Verification libraries (http://lib.stat.cmu.edu/R/CRAN/).

3. Results

The surgical excision samples were pathologically diagnosed as ADH or benign tumours in 81% of patients ($n = 87$) and as malignancy in 19% of patients ($n = 20$). Amongst the malignant cases, the histological results revealed the presence of invasive carcinoma and DCIS in 11% of cases ($n = 12$) and 8% of cases ($n = 8$), respectively. Table 1 summarises the underestimation rates according to the clinical, radiological and pathological factors. Important differences are noted between this validation dataset and the training set used by Ko et al.[15] Table 2 summarises the comparison based on the scoring system variables. The observed odds ratios (OR) and *P* values in

Table 1 – Pathological results after surgical excision according to clinical, radiological and pathological characteristics of 107 cases of atypical ductal hyperplasia (ADH).

Characteristics	Number of biopsies[a] (*n* (%))	Pathology after surgical excision		Underestimation rate (%)	P value (Chi-square)
		ADH (*n*)	Malignancy (*n*)		
Number of cases	107 (100)	87	20	18.7	
Age (years)					
⩽50	37 (35)	28	9	24	
>50	70 (65)	59	11	16	0.48
Palpability					
Impalpable lesion	106 (99)	86	20	19	
Palpable lesion	1 (1)	1	0	0	0.41
Menopausal status					
No	42 (39)	33	9	21	
Yes	64 (61)	54	10	16	0.61
Menopausal hormone therapy use					
No	89 (83)	73	16	18	
Yes	16 (17)	13	3	19	0.78
Personal history of breast carcinoma					
No	89 (87)	74	15	17	
Yes	13 (13)	10	3	23	0.87
Family history of breast carcinoma					
No	75 ()	65	10	13	
Yes	29 ()	20	9	31	0.07
Size of calcifications (cm)					
⩽1.5	56 (53)	47	9	16	
>1.5	51 (47)	40	11	22	0.63
BI-RADS					
C3	10(10)	8	2	20	
C4	88 (85)	73	15	17	
C5	5 (5)	2	3	60	–
Focal (ADH) (⩽1 duct and ⩽1 mm)	38 (35)	32	6	16	
Extensive ADH	69 (65)	55	14	20	0.75
Microcalcification on mammography					
No	10 (9)	9	1	10	
Yes	97 (91)	78	19	20	0.75

[a] The total number of biopsies was lower than 107 if the data were missing from the hospital records.

the multivariate analysis of these criteria were as follows: age >50 years, OR = 0.50 (95% CI, 0.18–1.45, $P = 0.21$); microcalcification on mammography, OR = 2.07 (95% CI, 0.23–18.32, $P = 0.51$); lesion size >1.5 cm, OR = 1.53 (95% CI, 0.52–4.45, $P = 0.43$); and focal ADH, OR = 1.79 (95% CI, 0.55–5.79, $P = 0.32$). The following additional variables (which were not included in the Ko nomogram) were not statistically significant: menopausal status, OR = 1.22 (95% CI, 0.17–8.61, $P = 0.83$); menopausal hormone therapy use, OR = 1.33 (95% CI, 0.22–7.90, $P = 0.75$); personal history of breast carcinoma, OR = 1.25 (95% CI, 0.25–6.27, $P = 0.77$); and family history of breast carcinoma, OR = 2.82. (95% CI, 0.90–8.89, $P = 0.07$).

3.1. *Performance of the logistic regression model in predicting invasive carcinoma*

Discrimination: The discrimination ability of the scoring system, which was measured using the area under the ROC curve (AUC), was 0.51 (95% CI, 0.47–0.53); this value reflected a poor discrimination performance.

Calibration: The calibration plot is shown in Fig. 2. The model is not well calibrated, and a significant difference was detected between the predicted and the observed probability ($P < 10^{-3}$). The average difference (E_{aver}) between the predicted and calibrated probabilities was 44%; this difference reflected a poor calibration performance.

3.2. *Performance of the scoring system in predicting invasive carcinoma (Table 3)*

Clinical utility: 18% of the samples received a score of ⩽3.5 and were defined as low-risk cases. The underestimation rate observed in this group was 11%. The performance of this threshold for sensitivity, specificity, positive predictive value, and negative predictive value was 90%, 22%, 20%, and 89%, respectively.

Table 2 – Comparison of the current and Ko population studies according to the scoring system criteria.

Variables	Current study (*n* (%))	Ko study (*n* (%))	*P* value (Chi-square)
Number of cases	107 (100)	74 (100)	
Age (years)			
⩽50	37 (35)	52 (70)	
>50	70 (65)	22 (30)	<0.001
Palpability			
Impalpable lesion	106 (99)	54 (73)	
Palpable lesion	1 (1)	20 (27)	<0.001
Size of calcifications (cm)			
⩽1.5	56 (53)	42 (57)	
>1.5	51 (47)	32 (43)	0.66
Focal atypical ductal hyperplasia (ADH) (⩽1 duct and ⩽1 mm)	38 (35)	8 (11)	
Extensive ADH	69 (65)	66 (89)	<0.001
Microcalcification on mammography			
No	10 (9)	42 (57)	<0.001
Yes	97 (91)	32 (43)	

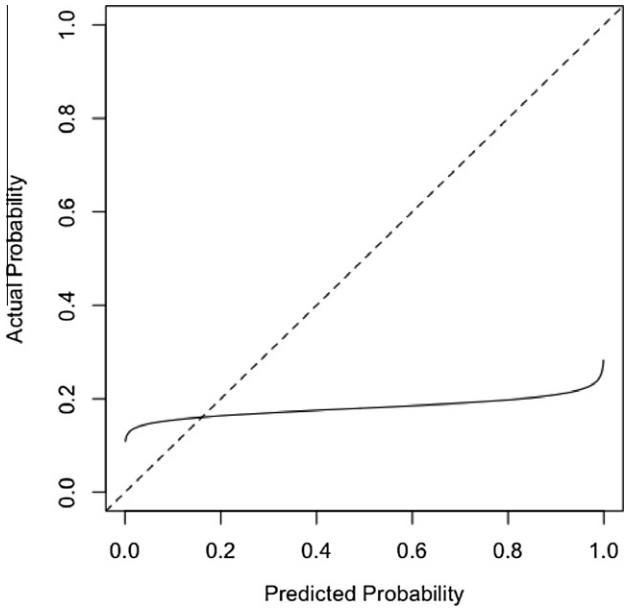

Fig. 2 – Calibration plot for the entire cohort of 107 patients. (*E*, difference in predicted and calibrated probabilities between the calibration and the area under the receiver operating characteristic curve; E_{aven}, average error = 44%; $P < 10^{-3}$. Ideal calibration line - - - -; Ko calibration line ___).

4. Discussion

External validation, which is more revealing than internal validation, is required to evaluate the performance and general applicability of a model before its implementation in clinical practice.[19–21] This study used validated statistical methods to assess the performance[16] of Ko's LRM and scoring system[15] in an independent multicentre population diagnosed with ADH by 11-gauge VABB. We observed an underestimation rate of 19% (20/109). Comparatively, Ko et al.[15] have reported an overall underestimation rate of 45.9% and 31.5% for the study (74 cases) and validation datasets (54 cases), respectively, from 11–14 vacuum-assisted CNB. The specific underestimation rate related to the 11-gauge VABB in the Ko et al. study was 37.5% (9/24). A variation between 4 and 33% is classically observed with 11-gauge VABB.[22–28] Several explanations for the low level of concordance (i.e. high rate of underestimation) between the pathology results from image-guided biopsy (IGB) and surgical excision can be proposed. The sampling method (needle gauge, number of cores obtained, use of vacuum assistance) for collecting small amounts of tissue causes the distinction between ADH and DCIS grade I in core biopsies to be problematic. Second, the high degree of inter-observer variability between pathologists due to the morphological continuum of intraductal proliferations from ADH to DCIS and the uncertainty related to the diagnostic thresholds can also contribute to exceedingly difficult and highly subjective distinctions. The use of a scoring system (integrating imaging data, history, etc.) is intended to overcome this difficulty and to select the patients that require surgery. Such a scoring system must indeed integrate both uncertainty from sampling and variability in the pathologic diagnosis (inter-observer variability); our study suggests that the criteria included in the Ko score are insufficient to define a subgroup with a less than 2% underestimation rate.

The multivariate logistic regression model and the scoring system were built using five variables, which can be separated into the following three categories: (i) clinical factors (age at biopsy and the palpability of the lesion), (ii) imaging factors (microcalcification on mammography and size of calcifications), and (iii) pathology results at biopsy (focal ADH). We observed a discrimination ability of the LRM using the AUC value of 0.51 (95% CI, 0.48–0.53) compared to 0.90 (95% CI, 0.83–0.97) in the Ko et al.[15] training set. A downgrade of the AUC value had already been observed between the training set and the validation dataset (0.85, 95% CI, 0.74–0.95) within the Ko study.[15] This trend suggested the possibility of overfitting.[29,30] The lack of accuracy could be due to the choice of predictors, which were selected only for their significant

Table 3 – Underestimation rate, sensitivity, specificity, and positive predictive and negative predictive values according to various scores.

Score	Benign (*n*)	Malignancy (*n*)	Underestimation rate (%)	Sensitivity	Specificity	PPV	NPV
⩽3.5	17	2	11	0.90	0.22	0.20	0.89
	70	18					
5.5–7.5	24	8	25	0.60	0.27	0.16	0.75
	63	12					
⩾9	46	10	18	0.50	0.53	0.20	0.82
	41	10					
Total	87	20	19	–	–	–	–

PPV: positive predictive value.
NPV: negative predictive value.

associations (based on *P* values) with histological underestimation in a multivariate analysis. Within our data, no factor was found to be independently predictive. In addition, the discrepancy between the studies could be explained by significant clinical, imaging and pathological differences between the two populations (Table 2). However, a robust nomogram must still have greater accuracy than that observed in this study. In theory, the performance of the nomogram (according to the discrimination criteria) must not be affected by these differences. Apart from the discrimination, which is limited by a poor clinical significance, we used the calibration measurements (Fig. 2) to provide better information regarding the true accuracy of the model. The scoring system was not well calibrated because the predicted percentages were unsatisfactory when both low- and high-risk patients were studied. In addition, we calculated the average error (E_{aver}) between the predictions and the observations obtained from the calibration curve. This value provides an idea of model performance when extrapolated to new patient populations. Our results indicated that this probability (44%) was not sufficiently accurate to be used for patient education. Aside from the calibration and discrimination, the clinical utility of the model may help clinicians. Ko et al.[15] have reported that a score ⩽3.5 allows a subset of patients to be designated as 'probably benign' and, thus, eligible for exclusively non-invasive follow-up. In the Ko study, the subset that corresponded to the clinical utility was 21.6% and 27.8% of the patients in the study and validation datasets, respectively.[15] In the current study, the largest low-risk subgroup observed was 18%, indicating that the Ko et al. scoring system assigned one fifth of the patients to the low-risk group. According to the conclusions of Ko et al., this subset should be eligible for non-invasive management based on sensitivity and negative predictive values of 100% and an underestimation rate of 0%. Thus, we applied this scoring to our population to test the score of ⩽3.5. Unfortunately, we were not able to reproduce these findings. We reported sensitivity and negative predictive values of 90% and 89%, respectively. The upgrading rate related to this score was 11%, indicating that two patients could not be well discriminated and this could lead to the possibility of undertreatment. This underestimation in clinical practice reflected the limits of the scoring system in external data. To conclude, using our data, the LRM was not sufficiently accurate for individual predictions of malignancy, and the scoring system was not able to safely define the subset of patients with a risk of less than 2%, which represents the group of patients who would be eligible for exclusively non-invasive follow-up. Therefore, our findings underline the lack of reproducibility and show the difficulty in the general application of the nomogram to another study group. The lack of accuracy and inability of this model and other published models, such as those of Gail[31] and Cusik,[32] to safely define a subgroup of patients who could be managed non-operatively emphasises the importance of a multidisciplinary approach. The application of some principles of multidisciplinary care correlated with clinicopathological and mammographic features could be the best practice recommendations to improve the management (diagnosis, treatment, and follow-up) of these patients. In addition, the lack of performance of these nomograms suggests the difficulty in predicting invasive cancer in ADH cases diagnosed by VABB. Other predictive methods must be actively investigated. Thus, improvements in predictions will likely be contingent upon further investigation of the molecular mechanism of ADH.

Conflict of interest statement

None declared.

REFERENCES

1. Dahlstrom JE, Sutton S, Jain S. Histological precision of stereotactic core biopsy in diagnosis of malignant and premalignant breast lesions. *Histopathology* 1996;**28**:537–41.
2. Parker SH, Lovin JD, Jobe WE, et al. Nonpalpable breast lesions: stereotactic automated large-core biopsies. *Radiology* 1991;**180**:403–7.
3. Ciatto SN, Houssami D, Ambrogetti S, et al. Accuracy and underestimation of malignancy of breast core needle biopsy: the Florence experience of over 4000 consecutive biopsies. *Breast Cancer Res Treat* 2007;**101**:291–7.
4. Jackman RJ, Birdwell RL, Ikeda DM. Atypical ductal hyperplasia: can some lesions be defined as probably benign after stereotactic 11-gauge vacuum-assisted biopsy, eliminating the recommendation for surgical excision? *Radiology* 2002;**224**:548–54.
5. Tavassoli FA, Devilee P. *WHO classification tumors of the breast and female genital organs*. WHO IARC; 2003.
6. Tavassoli FA, Norris HJ. A comparison of the results of long-term follow-up for atypical intraductal hyperplasia and intraductal hyperplasia of the breast. *Cancer* 1990;**65**:518–29.
7. Dupont WD, Page DL. Risk factors for breast cancer in women with proliferative breast disease. *N Engl J Med* 1985;**312**:146–51.

8. Rosen PP. Proliferative breast 'disease'. An unresolved diagnostic dilemma. *Cancer* 1993;**71**:3798–807.
9. Yeh IT, Dimitrov D, Otto P, et al. Pathologic review of atypical hyperplasia identified by image-guided breast needle core biopsy. Correlation with excision specimen. *Arch Pathol Lab Med* 2003;**127**:49–54.
10. Adrales G, Turk P, Wallace T, Bird R, Norton HJ, Greene F. Is surgical excision necessary for atypical ductal hyperplasia of the breast diagnosed by Mammotome? *Am J Surg* 2002;**180**:313–5.
11. Philpotts LE, Lee CH, Horvath LJ, et al. Underestimation of breast cancer with II-gauge vacuum suction biopsy. *AJR Am J Roentgenol* 2000;**175**:1047–50.
12. Forgeard C, Benchaib M, Guerin N, et al. Is surgical biopsy mandatory in case of atypical ductal hyperplasia on 11-gauge core needle biopsy? A retrospective study of 300 patients. *Am J Surg* 2008;**196**:339–45.
13. Graesslin O, Antoine M, Chopier J, et al. Histology after lumpectomy in women with epithelial atypia on stereotactic vacuum-assisted breast biopsy. *Eur J Surg Oncol* 2010;**36**:170–5.
14. American College of Radiology. *Breast imaging reporting and data system (BI-RADS)*. 4th ed. Reston, VA: American College of Radiology; 2003.
15. Ko E, Han W, Lee JW, et al. Scoring system for predicting malignancy in patients diagnosed with atypical ductal hyperplasia at ultrasound-guided core needle biopsy. *Breast Cancer Res Treat* 2008;**112**:189–95.
16. Coutant C, Olivier C, Lambaudie E, et al. Comparison of models to predict nonsentinel lymph node status in breast cancer patients with metastatic sentinel lymph nodes: a prospective multicenter study. *J Clin Oncol* 2009;**27**:2800–8.
17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.
18. Cox D. Two further applications of a model for binary regression. *Biometrika* 1958;**45**:562–5.
19. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;**130**:515–24.
20. McGinn TG, Guyatt GH, Wyer PC, et al. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. *JAMA* 2000;**284**:79–84.
21. Knottnerus JA. Prediction rules: statistical reproducibility and clinical similarity. *Med Decis Making* 1992;**12**:286–7.
22. Kohr JR, Eby PR, Allison KH, et al. Risk of upgrade of atypical ductal hyperplasia after stereotactic breast biopsy: effects of number of foci and complete removal of calcifications. *Radiology* 2010;**255**:723–30.
23. Allison KH, Eby PR, Kohr J, et al. Atypical ductal hyperplasia on vacuum-assisted breast biopsy: suspicion for ductal carcinoma in situ can stratify patients at high risk for upgrade. *Hum Pathol* 2010;**42**:41–50.
24. Penco S, Rizzo S, Bozzini AC, et al. Stereotactic vacuum-assisted breast biopsy is not a therapeutic procedure even when all mammographically found calcifications are removed: analysis of 4,086 procedures. *AJR Am J Roentgenol* 2010;**195**:1255–60.
25. Deshaies I, Provencher L, Jacob S, et al. Factors associated with upgrading to malignancy at surgery of atypical ductal hyperplasia diagnosed on core biopsy. *Breast* 2010;**20**:50–5.
26. Nguyen CV, Albarracin CT, Whitman GJ, et al. Atypical ductal hyperplasia in directional vacuum-assisted biopsy of breast microcalcifications: considerations for surgical excision. *Ann Surg Oncol* 2010;**18**:752–61.
27. Doren EM, Hulvat J, Norton P, et al. Predicting cancer on excision of atypical ductal hyperplasia. *Am J Surg* 2008;**195**:358–61.
28. Yu YH, Liang C, Yuan XZ. Diagnostic value of vacuum-assisted breast biopsy for breast carcinoma: a meta-analysis and systematic review. *Breast Cancer Res Treat* 2010;**120**:469–79.
29. Tetko IV, Livingstone DJ, Luik AI. Neural network studies. Comparison of overfitting. *J Chem Inf Comput Sci* 1995;**35**:826–33.
30. Degnim AC, Reynolds C, Pantvaidya G, et al. Nonsentinel node metastasis in breast cancer patients: assessment of an existing and a new predictive nomogram. *Am J Surg* 2005;**190**:543–50.
31. Boughey JC, Hartmann LC, Anderson SS, et al. Evaluation of the Tyrer-Cuzick (International Breast Cancer Intervention Study) model for breast cancer risk prediction in women with atypical hyperplasia. *J Clin Oncol* 2010;**28**:3591–6.
32. Pankratz VS, Hartmann LC, Degnim AC, et al. Assessment of the accuracy of the Gail model in women with atypical hyperplasia. *J Clin Oncol* 2008;**26**:5374–9.