

Reproducibility of mammographic classifications for non-palpable suspect lesions with microcalcifications

¹R M PIJNAPPEL, MD, PhD, ²P H M PEETERS, MD, PhD, ³J H C L HENDRIKS, MD, PhD and ⁴W P Th M MALI, MD, PhD

¹Department of Radiology, Martini Hospital, Locatie van Swieten, PO Box 30033, 9700 RM Groningen, ²Julius Centre for General Practice and Patient Oriented Research, Utrecht University Medical School, PO Box 85500, 3508 GA Utrecht, ³National Expert and Training Centre for Breast Cancer Screening, University Medical Centre Nijmegen, PO Box 9101, 6500 HB Nijmegen and ⁴Department of Radiology, University Medical Centre Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands

Abstract. Observer variability in interpretation of mammograms is a well-known problem, especially for microcalcifications. The classification of the mammographic findings depends upon this interpretation. We performed an intraobserver study to evaluate a breast imaging reporting and data system (BI-RADS) based method for description and classification of non-palpable lesions with microcalcifications. A set of 100 non-palpable mammographic lesions mainly consisting of microcalcifications was described and classified on two occasions, by two radiologists at an interval of 6 months. The intraobserver variability was evaluated with kappa statistics. The overall agreement for the classification was moderate (kappa 0.54). The lowest kappa values were observed for the categories “probably benign (BI-RADS 3, kappa 0.59)” and “suspicious abnormality (BI-RADS 4, kappa 0.44)”. The clinical management (follow-up or biopsy) of non-palpable lesions consisting of microcalcifications depending upon radiological classification in the groups BI-RADS 3 (follow-up) and BI-RADS 4 (biopsy) is therefore debatable.

The classification of non-palpable lesions consisting of microcalcifications on mammography in order to describe the radiologist's level of cancer suspicion is complex. This classification determines the final assessment of the lesion, histological biopsy or mammographic follow-up. One of the limitations of classifying mammographic lesions is the observer variability in interpretation of mammographic features [1, 2]. In order to produce a more standardized scheme for describing mammographic lesions, the American College of Radiology recommends the Breast Imaging Reporting and Data System (BI-RADS) [3]. This system includes terms for describing calcifications (morphology and distribution) and describes the radiologist's level of suspicion about the mammographic abnormality (Table 1). Several studies have evaluated this system [4, 5]. When BI-RADS is used, substantial interobserver and intraobserver variability in mammographic interpretation is noted [6]. Another limitation of the BI-RADS lexicon is that it does not make explicit recommendations regarding the assignment of final assessment categories for the different mammographic features [5].

We set out a study to evaluate the intraobserver variability of a BI-RADS based system for description and interpretation of non-palpable mammographic lesion mainly consisting of microcalcifications and analysed the clinical impact in terms of advising biopsy or follow-up.

Patients and methods

For this study we used data from the COBRA-study (core biopsy after radiological localization) [7]. This study evaluated the accuracy of stereotactic-guided large core

needle biopsy for impalpable lesions classified as non-benign. The study was conducted between April 1997 and February 2000. A total of 973 patients with 1029 non-palpable lesions classified as BI-RADS 3 to 5 from 19 Dutch hospitals were included. After informed consent, all patients were first offered a stereotactic-guided 14-gauge core needle biopsy. Whenever needle biopsy yielded a malignancy, therapeutic surgery was performed. In those cases where no malignancy was found on core biopsy, the patients underwent a needle localized breast biopsy. The final diagnosis was based upon the pathologic result of the surgical specimen (needle localized breast biopsy or mastectomy).

All impalpable lesions, mainly consisting of microcalcifications, in the COBRA study were identified ($n=533$) and copies of the original mammography were reviewed by two radiologists (RP, JH). Both are specialists in breast imaging. Lesions were described and classified in consensus. The microcalcifications could be classified according to a total of six morphology categories (linear, coarse granular, fine granular countable, very fine granular non-countable, “cottonwool” and coarse non-granular (benign impression)) [8]. All the different types of microcalcifications present in a lesion were classified. The morphology

Table 1. BI-RADS categories

| Category | Definition |
|----------|---|
| 1 | Normal mammogram |
| 2 | Benign finding |
| 3 | Probably benign finding; follow-up suggested |
| 4 | Suspicious abnormality; biopsy should be considered |
| 5 | Highly suggestive of malignancy; appropriate action should be taken |

Received 19 May 2003 and in revised form 13 October 2003, accepted 16 December 2003.

Table 2. Intraobserver variability in description of lesions with microcalcifications

| | Number first reading | Number second reading | Kappa value (SE) |
|--|----------------------|-----------------------|------------------|
| Microcalcifications morphology | | | |
| linear | 11 | 12 | 0.55 (0.11) |
| coarse granular | 51 | 54 | 0.58 (0.08) |
| fine granular countable | 29 | 28 | 0.43 (0.09) |
| very fine granular non-countable | 13 | 8 | 0.40 (0.12) |
| "cottonwool" | 6 | 7 | 0.63 (0.11) |
| coarse non granular | 12 | 13 | 0.58 (0.10) |
| Microcalcifications distribution (overall) | | | 0.49 (0.09) |
| distribution diffuse | 7 | 3 | 0.37 (0.19) |
| distribution in cluster(s) | 71 | 73 | 0.60 (0.09) |
| distribution segmental | 22 | 24 | 0.38 (0.11) |

SE, standard error.

and distribution of the microcalcifications were studied and lesions were classified into 3 groups (probably benign, suspicious and highly suggestive of malignancy) according to the BI-RADS lexicon 3 to 5. After a 6 month period a random sample of 100 out of these 533 mammograms were re-examined and classified again by the same team. We calculated the intraobserver variability and evaluated the description and interpretation with kappa statistics. We calculated the kappa values by comparing the results of the first and second reading session. Since one mammographic lesion may contain different morphological types of microcalcifications, we divided all morphological types into two groups, *i.e.* linear calcifications *vs* the rest; or cottonwool *vs* the rest, etc. By doing this we were able to calculate kappa values for each category. Perfect agreement is indicated by a kappa value of 1.0, whereas a kappa value of 0 indicates no agreement at all. Kappa is a percentage of agreement adjusted for chance fluctuations. Prior reports have suggested the following level of agreement between observers for the indicated kappa values: ≤ 0.20 , slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement

Table 3. Intraobserver variability of radiological classification

| | Kappa value (SE) |
|---|------------------|
| Radiological classification (overall) | 0.54 (0.07) |
| Probably benign (BI-RADS 3) | 0.59 (0.08) |
| Suspicious abnormality (BI-RADS 4) | 0.44 (0.09) |
| Highly suggestive of malignancy (BI-RADS 5) | 0.73 (0.09) |

BI-RADS, Breast imaging reporting and data system; SE, standard error.

Table 4. Inconsistencies of BI-RADS classification

| BI-RADS assessment first reading session | BI-RADS assessment second reading session | Number of inconsistencies | Malignant inconsistencies |
|--|---|---------------------------|---------------------------|
| 3 | 4 | 14/44 (32) | 2/14 (14) |
| 4 | 3 | 7/41 (17) | 3/7 (43) |
| 4 | 5 | 5/41 (12) | 4/5 (80) |
| 5 | 4 | 4/15 (27) | 2/4 (50) |

Numbers in parentheses are percentages.
SE, standard error.

between observers [4, 9]. Furthermore we studied the inconsistencies in the BI-RADS classification and correlated these cases with the final pathologic assessment.

Results

38 of the 100 lesions were malignant (*in situ* or invasive). The kappa values for calcification morphology and distribution are shown in Table 2. The agreement was slight for very fine granular non-countable microcalcifications (kappa=0.40). There was moderate agreement for description of the microcalcifications categories: linear, coarse granular, fine granular countable and coarse non-granular (kappa values 0.55, 0.58, 0.43 and 0.58, respectively). Only for cottonwool calcifications was the agreement substantial (kappa=0.63). The overall agreement for the distribution of microcalcifications was moderate (kappa 0.49).

Results of intraobserver variability in radiological classifications are shown in Table 3. For the BI-RADS categories 3 (kappa=0.59) and 4 (kappa=0.44) the agreement was moderate. For BI-RADS category 5 the agreement was substantial (kappa=0.73).

Table 4 shows the final histological results according to first and second classification of BI-RADS categories. During the first classification 44 lesions were classified as a BI-RADS 3, and 6 turned out to be malignant. 14 of these lesions were classified as BI-RADS 4 after the second classification. Two of these 14 lesions were malignant. 41 lesions were classified as a BI-RADS 4, 21 turned out to be malignant. Seven of these lesions were classified as BI-RADS 3 and 5 as BI-RADS 5 after the second session. Of these 7 and 5 differently classified lesions, 3 and 4 were malignant, respectively. 15 lesions were classified as BI-RADS 5 during the first classification, and 11 turned out to be malignant. Four of these lesions were classified as

BI-RADS 4 after the second classification. Two of these 4 were malignant.

Discussion

In general, there was only moderate agreement in describing microcalcifications according to morphology and distribution and thus in classifying the lesions radiologically according to BI-RADS. These results are in concordance with the literature. Previous studies showed slight to moderate overall intraobserver and interobserver agreement for the description of microcalcifications morphology (kappa varied from 0.36 to 0.57) [4, 6].

In our study the group of fine granular calcifications shows the least agreement. This might be due to dividing the whole group of fine granular calcifications into two subgroups (fine granular countable and fine granular non-countable). Dividing granular calcifications into subgroups appears to be of little value. Our result for describing the distribution of microcalcifications (kappa 0.47) is similar to a previous study on this subject [6].

Inconsistency was greatest for BI-RADS category 3 (probably benign) and 4 (suspicious). All radiologists involved with mammography know the limitations of interpretation and classification. For the referring physician the difference between BI-RADS 3 (follow-up) or BI-RADS 4 (biopsy procedure) is important. The moderate agreement in radiological classification for the categories 3 and 4 show that in cases where lesions with microcalcifications are non-benign but not highly suspicious of malignancy it is difficult to clearly categorize them. In the COBRA-study, as in most of the hospitals in the Netherlands, all non-benign lesions (including BI-RADS 3) underwent histological confirmation. Outside this study and strictly following the BI-RADS lexicon, lesions categorized as BI-RADS 3 might not have had a biopsy procedure but would have been evaluated by follow-up. If the BI-RADS guidelines for the management of lesions categorized as BI-RADS 3 were followed in our study the diagnosis of 6/38 (16%) malignancies would have been delayed following the first reading session, and 7/38 (18%) in the second reading session. Of these cases 2 appeared to be invasive carcinoma (one invasive ductal carcinoma and one invasive lobular carcinoma), 4 DCIS grade I and 5 DCIS grade II. This might have resulted in a worse prognosis for these patients. These results support the Dutch and the British strategy of obtaining a pathological diagnosis for lesions categorized as BI-RADS 3. We suggest that non-palpable lesions with microcalcifications categorized as BI-RADS 3 (probably benign) should undergo a biopsy procedure until a more reliable system for description and classification of microcalcifications is available. Indirect support for this strategy comes from another study that showed that malignancy was found in 24.8% of lesions where microcalcifications were stable and considered indeterminate or suspicious [10].

This study has several limitations. First, all lesions were classified as non-benign by a radiologist before entering this study; therefore in this study only selected cases were used. Second, the two observers in this study work closely together and one observer (JH) was responsible for the training of the other one (RP). Therefore our results may vary somewhat from results for radiologists at other institutions. Third, all descriptions and classifications were made in consensus, there was no individual observer description. Fourth, all mammograms were copies, introducing loss of image quality. Nevertheless, overall we found only moderate agreement.

In conclusion, the agreement in classification of non-palpable lesions with microcalcifications categories as BI-RADS 3 and BI-RADS 4 is moderate. Therefore we think that for non-palpable lesions with microcalcifications classified as non-benign (higher as BI-RADS 2) it is advisable to advocate a biopsy procedure. Trying to classify those lesions in the groups probably benign (BI-RADS 3) or suspicious abnormality (BI-RADS 4) has limited clinical value.

References

1. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med* 1994;331:1493-9.
2. Kopans DB. Accuracy of mammographic interpretation. *N Engl J Med* 1994;331:1521-2.
3. American College of Radiology. Breast imaging reporting and data system (BI-RADS), 2nd edn. Reston, VA: American College of Radiology, 1995.
4. Baker JA, Kornguth PJ, Floyd CE. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. *AJR Am J Roentgenol* 1996;166:773-8.
5. Liberman L, Abramson AF, Squires FB, Glassman JR, Morris EA, Dershaw DD. The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories. *AJR Am J Roentgenol* 1998;171:35-40.
6. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast imaging reporting and data system: inter- and intraobserver variability in feature analysis and final assessment. *AJR Am J Roentgenol* 2000;174:1769-77.
7. Verkooijen HM. Diagnostic accuracy of stereotactic large-core needle biopsy for nonpalpable breast disease: results of a multicenter prospective study with 95% surgical confirmation. *Int J Cancer* 2002;99:853-9.
8. Tabar L, Dean P. Teaching atlas of mammography (2nd edn). Stuttgart: Georg Thieme Verlag, 1985.
9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
10. Lev-Toaff AS, Feig SA, Saitas VL, Finkel GC, Schwartz GF. Stability of malignant breast microcalcifications. *Radiology* 1994;192:153-6.