Histologic Grading in Breast Cancer

Reproducibility Between Seven Pathologic Departments

Poul Boiesen, Pär-Ola Bendahl, Lola Anagnostaki, Henryk Domanski, Erik Holm, Ingrid Idvall, Sven Johansson, Otto Ljungberg, Anita Ringberg, Görel Östberg and Mårten Fernö for the South Sweden Breast Cancer Group

From the Pathologic and Cytologic Departments at the Hospitals in Helsingborg (P. Boiesen), Malmö (L. Anagnostaki), Kristianstad (H. Domanski, O. Ljungberg), Karlskrona (E. Holm), Lund (I. Idvall), Växjö (S. Johansson), Halmstad (G. Östberg), the Department of Oncology, University Hospital in Lund (P.-O. Bendahl, M. Fernö), and the Department of Plastic Surgery, University Hospital in Malmö (A. Ringberg), Sweden

Correspondence to: Dr Mårten Fernö, Department of Oncology, University Hospital, S-221 85 Lund, Sweden. Tel: + 46 46 17 75 65. Fax: + 46 46 14 73 27. Email: marten.ferno@onk.lu.se

Acta Oncologica Vol. 39, No. 1, pp. 41–45, 2000

Histologic grade, including tubular formations, nuclear grade, and mitotic activity, is a well-documented prognostic factor in breast cancer. In comparison with other prognostic parameters, the evaluation of histologic grade is cheap and can be performed, in principle, in all cases of breast cancer. One possible disadvantage is that the evaluation may vary between different pathological departments. The aim of the present work was therefore to study the reproducibility of the histologic grading system by distributing haematoxylin-erythrosin-stained slides from 93 invasive breast cancers to the seven pathology departments within the southern healthcare region of Sweden. The evaluation was performed blindly and without any knowledge of other clinical parameters. In 31% of the cases the same histologic grade was obtained for all departments. The overall mean kappa was 0.54, indicating a moderate reproducibility. Of the three factors included in histologic grade, the agreement was best for tubular formations and poorest for nuclear grade and mitotic activity. The overall moderate reproducibility should be considered when the clinical usefulness of histologic grading is compared with other prognostic instruments.

Received 22 February 1999
Accepted 18 May 1999

At the time of primary operation of breast cancer patients it is desirable to be able to predict the clinical course of the disease. Hitherto, it has been generally accepted that the TNM classification (tumour size, lymph node status, and presence or absence of metastases) provides clinically useful prognostic information. As the fraction of early-detected small breast cancers without lymph node involvement has increased in the past decade, additional prognostic factors are needed. Histologic grade, steroid receptors, and markers of proliferation are examples of factors giving clinically useful prognostic information. Besides providing prognostic information, a clinically useful factor should be measurable for as many tumour samples as possible, using an easy, cheap, and reproducible technique. Evaluation of histologic grading on paraffin-embedded sections has most of these advantages, but the interobserver reproducibility of histopathological features has been claimed to be unsatisfactory (1). When a histologic grading scheme with specified guidelines was used, however, the reproducibility was found to be acceptable (2–4). One commonly applied system is the Nottingham Prognostic Index (NPI), which is based on histologic grade, tumour size, and lymph node status (5). The histologic grade used in this index is a modification of the Bloom and Richardson system (6), and is based on tubular formations, nuclear grade, and mitotic activity (5). In several studies, the NPI has been found to give valuable prognostic information in breast cancer, and because of its simplicity and availability, it is considered to be suitable for routine clinical use (7–9). But before being used as a clinical routine, the NPI should be compared with other prognostic instruments with regard to reproducibility and prognostic strength. The aim of the present study was to evaluate the reproducibility of histologic grading for experienced pathologists at the seven pathology departments in the southern healthcare region of Sweden.

MATERIAL AND METHODS

Study design

Round 1. A re-evaluation of 108 slides, stained at the

 ISSN 0284-186X

primary histological evaluation, from breast cancer patients with stages I and II disease, was carried out at seven pathology departments in the southern healthcare region of Sweden. The patients were operated on between 1 January 1994 and 31 March 1995 and referred to the Department of Oncology, University Hospital in Lund for postoperative radiotherapy. The study was designed in such a way that the evaluation by each pathologist was performed blindly without any knowledge of patient and clinical characteristics, or of the judgements of the other pathologists. Fifteen slides were excluded from further analysis because of poor staining quality, making them impossible to evaluate, or because of no invasive cancer on that particular slide. The remaining 93 cases were considered by all pathologists to be invasive carcinoma, and therefore also included in the reproducibility evaluation.

Round 2. Six months after the first round, a representative subset (n = 31) of the 93 cases was distributed to the seven pathologic departments for a second evaluation. This subset had the same kappa value as the whole series (in the first round). The pathologists were not aware of which samples from the first round were sent out again.

Histopathological staining

All cases were routinely prepared, fixed in 10% buffered formalin, embedded in paraffin, sectioned at 4–5 μm and stained with haematoxylin-erythrosin. Histopathological evaluation of the pertinent parameters was performed blindly by the breast pathologist at each department. The histologic grade was obtained by adding the scores of tubular formations, nuclear grade, and mitotic activity, each parameter ranging from 1 to 3. Histologic grade 1 includes scores of 3–5, histologic grade 2 scores of 6 and 7, and histologic grade 3 scores of 8 and 9 (5).

Statistics

A generalized kappa (10) was used as a measure of agreement between the ratings of the seven pathologic departments. This measure, which is appropriate for more than two ratings and a constant number of rates, is often well approximated by the unweighted mean of all pairwise kappa values. In this study these values both happened to be 0.54. Two slightly different definitions of kappa were used throughout the paper—the generalized kappa, which is referred to as kappa or overall kappa, and a mean kappa per department, which is an average of pairwise kappa values between one department and the other six departments. Landis & Koch (11) give the following guidelines (it should, however, be emphasized that these guidelines are not general rules, since in some situations everything but a kappa equal to one is unacceptable, and thus the interpretation of kappa must be problem-specific).

Value of kappa	Strength of agreement
<0.20	Poor
0.20–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good

The subset of samples used in the second part of the study was randomly chosen, but with one restriction. Samples were drawn until a subset with the same generalized kappa as in the first part was found. The appropriate size of the subset was determined as follows: First, bootstrap was used to find estimates of the standard error of the generalized kappa for different sample sizes. Then the size was chosen so that the chance of detecting an improvement from kappa = 0.54 to kappa = 0.61, which is often referred to as the lower limit for 'good agreement', was about 80% using a one-sided test at the 5% level. A sample size of about 30 was found suitable. The sample size actually used (n = 31) is exactly one-third of the sample size in part one.

χ^2-test with 12 degrees of freedom was used to test differences in the distribution of ratings between departments. The software package Stata 5.0 was used for statistical analyses (12).

Table 1

The results of the first round, with the corresponding results from the second round shown in parentheses

	Histologic grade		Tubular formations		Nuclear grade		Mitotic activity	
Overall mean kappa	0.54	(0.50)	0.61	(0.54)	0.44	(0.29)	0.46	(0.31)
Complete agreement	31%	(26%)	45%	(42%)	24%	(3%)	33%	(10%)
Discrepant result by 0–1 departments	56%	(52%)	70%	(68%)	53%	(45%)	49%	(45%)
'Correctly' classified[1]	82%	(84%)	84%	(86%)	80%	(71%)	78%	(72%)
Disagreement 1–3	5/93	(3/31)	2/93	(2/31)	5/93	(4/31)	22/93	(14/31)
Best/worst concordance (%)[2]	77/57	(81/48)	86/68	(84/58)	66/57	(81/32)	67/56	(77/19)

[1] Assuming that the majority was right
[2] Between two departments

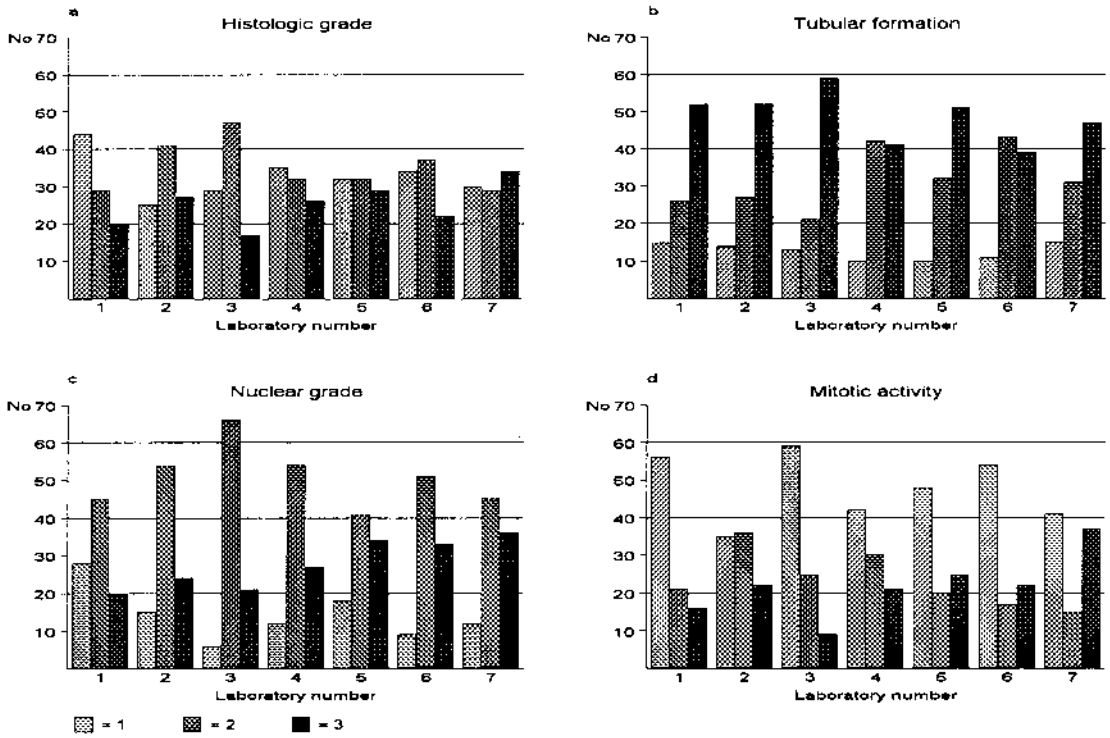

Fig. 1. (a)–(d) The distribution of histologic grade (1–3), tubular formations (1–3), nuclear grade (1–3), and mitotic activity (1–3) when the seven pathology departments in the southern healthcare region of Sweden evaluated the same 93 haematoxylin-erythrosin-stained slides.

RESULTS

Round 1. Histologic grade. Agreement between all departments was complete for 29 (31%) of the 93 samples (Table 1). Discrepant results were reported by 0–1 department (not necessarily the same one) for 52 (56%) of the samples, and by 0–2 departments for 73 (78%) of the samples. The best and worst concordance between two departments was 77% and 57%, respectively. For five of the samples all three histologic grades were reported. The department manifesting the best concordance with the others had a mean kappa value of 0.58 (range of six pairwise comparisons: 0.49–0.66), compared with 0.47 for the department manifesting the worst concordance (range: 0.39–0.60). The overall mean kappa was 0.54. The distribution of histologic grade differed systematically between the seven departments (see Fig. 1 (a); ($p = 0.034$)), with the following variations in histological grades between the departments: 27–47% (grade 1), 31–51% (grade 2), and 18–37% (grade 3). If it was considered that the majority was right for each sample, histologic grade was 'correctly' classified in 534 (82%) of a total 651 (93×7) judgements. The percentage of 'correctly' classified cases varied between 73% and 86% for the departments.

Tubular formations, nuclear grade, and mitotic activity

The distribution of tubular formations differed less than nuclear grade and mitotic activity between the seven departments ($p = 0.050$ (see Fig. 1 (b), tubular formations), $p < 0.001$ (Fig. 1 (c), nuclear grade), and $p < 0.001$ (Fig. 1 (d), mitotic activity)). The higher degree of agreement for tubular formations is also shown in Table 1, with an overall mean kappa of 0.61 compared to 0.44 for nuclear grade and 0.46 for mitotic activity. For tubular formations, the highest mean kappa value for a department was 0.64 (range of pairwise comparisons: 0.52–0.76), compared with 0.55 (range: 0.45–0.65) for the department manifesting the worst concordance. The corresponding figures for nuclear grade were 0.50 (0.32–0.57) and 0.32 (0.27–0.39), and for mitotic activity 0.52 (0.42–0.62) and 0.39 (0.30–0.50). The highest percentage of 'correctly' classified samples for one department was also higher for tubular formations than for nuclear grade and mitotic activity (91% vs. 86% and 82%, respectively; Table 1).

Round 2. Although the kappa values of the second round ($n = 31$) were somewhat lower than the those of the first round, the overall results for the two rounds were quite similar (Table 1). As the same pathologist examined the same slide twice, it was also possible to compare the intra-department reproducibility (Table 2). The re-evaluation resulted in altered histologic grade in 49 (23%) of the total 217 (7×31) evaluations. In 32 of the 49 disagreements the evaluation changed from

grade 3 to 2 or vice versa. The disagreement in histologic grade varied from 6% to 35% for the seven departments. The corresponding variations for the three components of histologic grade were: 6–19% (tubular formations), 16–39% (nuclear grade), and 23–52% (mitotic activity).

DISCUSSION

This study, concerning interobserver variation of histologic grade (tubular formations, nuclear grade, and mitotic activity) between seven pathological departments, has demonstrated a moderate reproducibility with an overall mean kappa value of 0.54. Since these results have been obtained with only a minimum of co-education, it should be possible to achieve an improvement of the reproducibility. However, a second round was performed after 2 co-education sessions, without any improvement in the reproducibility being found. The present study demonstrates that the concordance in the judgement of tubular formations was better than that of nuclear grade and mitotic activity. For 22 of the 93 cases all 3 mitotic classes were reported, compared to 5 for nuclear grade and histologic grade, and 2 for tubular formations. The reason for this difference is that the judgement of mitotic activity is impaired by several problems: poor quality of the slides with fixation artefacts, properties of the tumour such as necrosis, severe inflammation or extensive fibrosis. Furthermore, technical problems in deciding the size of the microscopic field or the finding of 'hot spots' of mitosis may cause difficulties. The results obtained in our study were similar to those reported in the investigation by Frierson and colleagues, in which six surgical pathologists evaluated histologic grade and each of its components for 75 infiltrating ductal carcinomas (3). They obtained generalized kappa values for histologic grade, tubular formations, nuclear pleomorphism, and mitotic count of 0.55, 0.64, 0.40, and 0.52, respectively, values that are very close to ours (0.54, 0.61, 0.44, and 0.46). The reproducibility for scoring nuclear pleomorphism in their work was clearly inferior to that for grade as well as for each of the other components. Frierson et al. suggested that using image analysis for evaluation of nuclear size and contour chromation distribution might in the future lead to a more accurate assessment of nuclear pleomorphism. In two other studies, higher kappa values were obtained. Dalton and co-workers (2) obtained a median weighted kappa value of 0.70 when 10 cases were evaluated according to the modified criteria of the Bloom and Richardson system (13). They considered that the cases with low and high scores in this system were reliable, whereas cases with an intermediate score should be evaluated by more than one pathologist. It is also in this intermediate group that other prognostic modalities may show the greatest promise (2, 14). Robbins and colleagues reported kappa values of 0.73 and 0.58, when 50 consecutive cases of breast cancer were fixed in both B5 (a mercuric chloride formalin mixture) and buffered formalin saline and evaluated at two hospitals by consensus of two and three pathologists, respectively (4). In this study the importance of fixation was suggested. In four other studies comparing levels of interobserver agreement for grading of breast cancer, kappa values of 0.17, 0.29, 0.57, and 0.66 were obtained (15–18).

When comparing the results from the first with those from the second round, it was also possible to evaluate the intra-department reproducibility. Almost one-quarter of the evaluations were altered and for 15% of the cases there was a change either from histologic grade 2 to histologic grade 3 or vice versa. The number of altered evaluations varied considerably between the departments. One department reported 35% discordant judgements of histologic grading, which should be compared with 6% for the department with the best concordance. It should also be mentioned that for one department all nine changes turned to a more aggressive grade, whereas for another department all three changes turned to a less aggressive grade. For a third department, all of the 11 altered evaluations were judged as grade 2 (6 turned from 3 to 2 and 5 from 1 to 2). Although there was no overall systematic pattern

Table 2

Number of discordances between Round 1 and Round 2 for each of the seven pathologic departments out of a total 31 comparisons

Department	Histologic grade	Tubular formations	Nuclear grade	Mitotic activity
1	11	5	12	11
2	7	6	7	8
3	2	6	5	7
4	9	6	9	16
5	3 #	2	11	6 #
6	9	4	12	9
7	7	6	7	10
Total number of disagreements	48	35	63	67

In one case mitotic activity, and consequently also histologic grade, was not considered possible to evaluate in Round 2.

for the changes, one cannot exclude that the 2 co-education sessions between the two evaluation rounds may have influenced the pathologists. This comparison furthermore revealed that a better agreement is more readily obtained for tubular formations than for nuclear grade and mitotic activity.

To sum up: the overall moderate reproducibility of histologic grade should be considered when its clinical usefulness is compared with other prognostic instruments. Furthermore, this study emphasizes the need for co-education for the evaluation of histologic grade and the use of standardized protocols.

ACKNOWLEDGEMENTS

This work was supported by grants from the Swedish Cancer Society, the Gunnar, Arvid and Elisabeth Nilsson Foundation, the Berta Kamprad Foundation, and the University Hospital of Lund Research Foundations.

REFERENCES

1. Gilchrist KW, Kalish L, Gould VE, et al. Interobserver reproducibility of histopathological features in stage II breast cancer. An ECOG study. Breast Cancer Res Treat 1985; 5: 3–10.
2. Dalton LW, Page DL, Dupont WD. Histologic grading of breast carcinoma. A reproducibility study. Cancer 1994; 73: 2765–70.
3. Frierson HF, Wolber RA, Berean KW, et al. Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. Am J Clin Pathol 1995; 103: 195–8.
4. Robins P, Pinder S, de Klerk N, et al. Histological grading of breast carcinomas: a study of interobserver agreement. Hum Pathol 1995; 26: 873–9.
5. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat 1992; 22: 207–19.
6. Bloom HJG, Richardson WW. Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer 1957; 11: 359–77.
7. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 1991; 19: 403–10.
8. Balslev I, Axelsson CK, Zedeler K, Bruun Rasmussen B, Carstensen B, Mouridsen HT. The Nottingham Prognostic Index applied to 9 149 patients from the studies of the Danish Breast Cancer Cooperative Group (DBCG). Breast Cancer Res Treat 1994; 32: 281–90.
9. Sauerbrei W, Hübner K, Schmoor C, Schumacher M, for the German Breast Cancer Study Group. Validation of existing and development of new prognostic classification schemes in node negative breast cancer. Breast Cancer Res Treat 1997; 42: 149–63.
10. Fleiss JL. Statistical methods for rates and proportions 2nd ed. New York: John Wiley & Sons, 1981: 230.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159–74.
12. StataCorp. Stata Statistical Software: Release 5.0 College, Station: TX, Stata Corporation, 1997.
13. Elston CW. Grading of invasive carcinoma of the breast. In: Page DL, Anderson TJ, eds. Diagnostic histopathology of the breast. Edinburgh: Churchill Livingstone, 1987: 300–11.
14. Barnes DM, Dublin EA, Fisher CJ, Levison DA, Millis RR. Immunohistochemical detection of p53 protein in mammary carcinoma: an important new independent indicator of prognosis? Hum Pathol 1993; 24: 469–76.
15. Davis BW, Gelber RD, Goldhirsch A, et al. Prognostic significance of tumor grade in clinical trials of adjuvant therapy for breast cancer with axillary lymph node metastasis. Cancer 1986; 58: 2662–70.
16. Hopton DS, Thorogood J, Clayden AD, MacKinnon D, for the Yorkshire Breast Cancer Group. Observer variation in histological grading of breast cancer. Eur J Surg Oncol 1989; 15: 21–3.
17. Theissig F, Kunze KD, Haroske G, Meyer W. Histological grading of breast cancer. Interobserver, reproducibility and prognostic significance. Path Res Pract 1990; 186: 732–6.
18. Harvey JM, de Klerk NH, Sterrett GF. Histological grading in breast cancer: interobserver agreement, and relation to other prognostic factors including ploidy. Pathology 1992; 24: 63–8.