J Clin Pathol 1983;**36**:392–398

Histopathological systems of breast cancer classification: reproducibility and clinical significance

B STENKVIST, E BENGTSSON, O ERIKSSON, T JARKRANS, B NORDIN, S WESTMAN-NAESER

From the Department of Clinical Cytology, University Hospital, S-751 85 Uppsala, Sweden

SUMMARY The inter- and intraobserver reproducibilities of the histopathological systems of breast cancer classification suggested by the World Health Organisation (WHO), the Armed Forces Institute of Pathology (AFIP) and Ackerman have been analysed. The reproducibilities of the three classification systems were only "fair" to "moderate" and no correlation with the five-year recurrence rate was found. Our results indicate that these classification systems are without biological significance and are useless for prognosis in the individual patient.

When the tumours were classified according to degree of differentiation (high, moderate, low) or graded according to WHO (which includes both differentiation and nuclear atypia), however, there was a significant correlation with the five-year recurrence rate. Yet even such "reduced" subdivisions are of no value in judging prognosis for the individual patient at the time of diagnosis; rather, they are useful only in the follow-up analysis of groups of patients.

The various systems for histopathological classification of breast cancers in use at present are descriptive, based on histological or cytological appearance, or both, of the tumours. While it has been claimed that tumour histology is of significance in some cases, in that some histological types behave less aggressively than others,[1 2] the usefulness of the different classification systems in evaluating individual patient prognosis and in selecting treatment has been questioned.[3]

This report, part of an extensive study of the morphological and epidemiological characteristics of breast carcinoma, analyses the reproducibility and biological significance of three well known systems of breast cancer classification—namely those of the WHO,[4] the AFIP[5] and Ackerman.[6] We have used Stewart's original classification (AFIP) and not the somewhat modified classification proposed later by McDivitt, Stewart and Berg.[7] After this study had begun, yet another classification system was suggested.[8] However, since the separate components forming the basis of the histopathological classification systems were analysed with respect to inter- and intraobserver reproducibility, we do not think that a separate analysis of these two later modifications would have influenced our conclusions.

Accepted for publication 23 November 1982

Material and methods

MATERIAL

Material was collected from 175 of 181 consecutively diagnosed breast cancer cases in four Swedish counties during the five-month observation period. Four of the cases were inoperable at diagnosis, and two patients refused treatment.

The cancers were unilateral and none had been given preoperative radiotherapy. The methods of operation varied. Either a simple mastectomy, radical mastectomy or mastectomy with exploration or exaeresis of the axilla was performed. After the operation for mammary carcinoma the patients were regularly checked at their local hospitals. Information concerning the occurrence of metastases or death was obtained from hospital records. The patients were followed up for five years.

TYPES OF SPECIMENS

Material for histopathological examination was obtained from all the patients subjected to surgical intervention—that is, exploratory biopsy or mastectomy. On removal all specimens were marked by the surgeon at the 12 o'clock position. Immediately after removal, the tumour was cut through and the two longest, perpendicular diameters were measured.

A slice (approx 3 mm thick) of the tumour and its

immediate surroundings was cut through its largest diameter and fixed in Carnoy's solution for 24 h. The rest of the specimen was fixed in 4% formaldehyde and sent to Uppsala, where all further processing of the material was performed.

The histopathological examination was performed by two of us (BS, SWN) according to a predesigned schedule and the data recorded for computer analysis. The observations were performed independently—twice—with an interval of 4–6 months—by each pathologist, in order to determine the degree of intra- and interobserver reproducibility. All histopathological studies were completed without prior knowledge of the clinical course and diagnosis.

Macroscopic tumour characteristics were noted, such as the weight of specimen, localisation of the tumour, distance between the centre of the tumour and the nipple, and involvement of the deep resection margin. The appearance and number of axillary lymph nodes was recorded.

Blocks for embedding in paraffin were taken from the tumour, the skin overlying the tumour, the deep line of excision, the nipple, the lymph nodes, one central and two peripheral parts of each of the four quadrants of the breast.

Sections (4 μm) were cut from the material fixed in formalin and Carnoy's solution. Haematoxylin and eosin and van Gieson's stain were used routinely. The following special stains were applied to the specimens fixed in Carnoy's solution and to lymph nodes with metastases: PAS stain with and without antecedent diastase digestion for the demonstration of glycogen and mucin, an elastic-fibre stain with orcein-iron haematoxylin according to Voerhoff, a connective-tissue stain according to Azan-Heidenhain, and methyl green-pyronine. This last stain was found to be valuable in identifying mast cells as well as plasma cells. The presence of keratin and mucus were also estimated in slides after staining with alcian green. The microscopic investigation was based on at least eight slides from each primary tumour or lymph node with metastatic growth of cancer.

Principles of tumour classification

The principles of classification in the three systems are similar in many respects. Tumours were classified according to the following scheme:

Common to all three systems are such concepts as "non-invasive", "mucinous" and "medullary carcinoma." In the classification systems of WHO and AFIP the concept "papillary carcinoma" is also used, whereas in Ackerman's classification system this concept seems to have its place in group II:3 ("well differentiated adenocarcinomas").

LOBULAR CARCINOMA

This is a specific entity in the WHO and the AFIP classification systems. The concept of lobular carcinoma is not explicitly included as a particular category of Ackerman's system.

OTHER (RARE) TUMOURS

The WHO and AFIP systems make room for all kinds of rare tumours, such as carcinosarcomas, acinic-cell, adenoid-cystic, epidermoid, spindle-cell, sweat-gland carcinomas, etc, whereas Ackerman's system does not seem to include such rare entities. Nor does Ackerman include such tumours as haemangiosarcomas, fibrosarcomas and lymphoma. None of the systems treat tubular carcinoma[9] as a separate entity.

INFILTRATING CARCINOMA

When the above mentioned tumours have been identified microscopically, all three systems include a broad variety of infiltrating carcinomas. In the WHO system, this class is referred to as "infiltrating carcinoma (class II)", whereas the AFIP system calls this group "carcinoma with fibrosis." Ackerman includes these carcinomas in his group "adenocarcinoma" (III:1) or "intraductal carcinoma with stromal invasion" (III:2).

HIGHLY METASTASISING CARCINOMA

As a subgroup of the infiltrating carcinomas, Ackerman introduced an additional group entitled "highly metastasising carcinomas." This is based on the classification of Hultborn and Törnberg,[10] which is, in its turn, a modication of Stewart's (AFIP) classification. Essentially, this group includes tumours with low degrees of differentiation or a diffuse invasion of the fat surrounding the tumour and an absence or scarce representation of lymphocytic response or vascular invasion of tumour cells.

Differentiation of breast carcinoma

In addition to the WHO, AFIP and Ackerman classification systems, we have introduced into this study the degree of differentiation as a separate entity. We scored carcinomas as having high, medium or low degrees of differentiation if the tumour was classified as invasive and operable. That is, non-invasive or non-epithelial or non-operable tumours were not so scored.

Differentiation was chosen as a more general concept than "tubule formation" of the WHO gradation system, since "tubule formation" is only one expres-

sion of differentiation. Other well differentiated epithelial tumours could thus be scored based on their abilities to form lobules or acini or ducts.

GRADING OF BREAST CARCINOMAS

The WHO system adds to tumour classification a system of grading based on a suggestion by Bloom and Richardson.[11] This is based on the number of hyperchromatic nuclei and mitoses per high power field of vision (1, 2 or 3 points: few, moderate, many), irregularity of size, shape and staining of nuclei (1, 2 or 3 points) and tubule formation (1, 2 or 3 points: many, moderate, none). These points are added together and 3–5 = grade I; 6–7 = grade II; 8–9 = grade III. No account is given either in the WHO book or in Bloom and Richardson's paper[11] of why these three variables in particular are used and what is more important, why the three characteristics are given the same weight in the gradation. Moreover, no account is given of how the regression analysis was performed. We have recently published an analysis of the reproducibility of the WHO, Black and Hartveit gradation systems.[12]

STATISTICAL METHODS

One method of measuring the inter- and intraobserver variabilities is to compute the percentage of equal assessments. However, this figure depends on the number of alternatives and if one category is chosen frequently in two judgements to be compared, the percentage of equal judgements will be high, even if there is statistical independence. Thus, a low value indicates poor agreement, but a high value may be misleading.

In our study, most of the information to be analysed was nominal, which makes the material suitable for correlation-coefficient analysis according to Tschuprow,[13] as well as by the percentage of equal assessments.

Another method of analysing the reproducibility used a coefficient of agreement of nominal scales, as described by Cohen.[14] On the basis of the proportion in which the judges agreed (p_o) and the proportion of units for which agreement is expected by chance (p_c), Cohen's kappa was calculated according to the formula

$$\text{Cohen's } \kappa = \frac{p_o - p_c}{1 - p_c}$$

This formula expresses the disagreements expected by chance. An approximation of the standard error of κ is given by the formula

$$\kappa = \sqrt{\frac{p_o(1 - p_o)}{N(1 - p_c)}}$$

Landis and Koch[15] have divided the relative strength of agreement associated with kappa statistics in order to provide "benchmarks" for discussion, as follows:

Kappa statistic	*Strength of agreement*
0·00	Poor
0·01–0·20	Slight
0·21–0·40	Fair
0·41–0·60	Moderate
0·61–0·80	Substantial
0·81–1·00	Almost perfect

SCORES OF THE VARIABLES

In order to determine a "final" class for each of the 174 tumours representing a "summing-up" classification of each tumour (before we correlated tumour class and recurrence rate) after the four readings that were performed, we proceeded as follows:

When three or four readings were in agreement, it was obvious that the tumour should be referred to

Table 1

Specimen	*Class*					*Score*
	I	*II*	*III*	*IV*	*V*	*(strictly final)*
1	a1b1		a2		b2	Class I
2		a1b1 a2	b2	b2		Class II
3		a1b1		a2		Class II
4	a1	a2			b2b1	Class V
						Approximated final score (see text about "winning reading")
5	a1 b1	a2 b2				Class I
6	a1		a2	b1	b2	Class IV

Table 2 *Inter- and intraobserver reproducibilities of breast cancer classification according to the WHO, the AFIP and Ackerman*

	WHO			*AFIP*			*Ackerman*		
	Inter	*Intra 1*	*Intra 2*	*Inter*	*Intra 1*	*Intra 2*	*Inter*	*Intra 1*	*Intra 2*
Cohen's κ	0·49	0·47	0·38	0·46	0·40	0·35	0·24	0·45	0·32
p value of χ^2 test	NA	NA	NA	NA	NA	NA	0·02	0·01	0·03
Tschuprow's coefficient	NA	NA	NA	NA	NA	NA	0·221	0·311	0·268
% equal assessments (No of classified tumours)	78·6 (168)	75·7 (173)	69·2 (169)	73·8 (168)	64·0 (172)	63·9 (166)	45·5 (165)	58·7 (167)	57·0 (165)

NA = not applicable.

the particular class chosen in the majority of the readings.

Table 1 illustrates how the scoring was performed when there was a greater spread of diagnoses at different readings (a_1; a_2; b_1; b_2). As an example of the problem in some cases in which a tumour was referred to a particular class, let us consider specimen 6 in Table 1. This specimen was referred to different classes at all four readings. (There is an analogous problem in the case of specimen 5, in which two readings gave one tumour class and two another). Since b_1 has occurred among the "winning" scores four times (specimens 1, 2, 3, and 4), a_1 three times (specimens 1, 2 and 3), a_2 once (specimen 2) and b_2 once (specimen 4), specimen 6 was scored as class 4—that is, the reading number 1 of pathologist b, since that reading most frequently belonged to the "majority opinion" in the classification of specimens 1–4.

In instances in which no "winning" reading was recorded, we classified the tumour by randomisation.

The classification procedure had a somewhat peculiar effect in a few instances in that a tumour could be classified differently by each system despite the fact that both systems used the same concept—that is, a tumour was classed as a "mucinous carcinoma" in one classification system and as an unspecified "infiltrating carcinoma" in another.

The procedure described above was followed for all the histopathological variables, in order to determine a final score for each variable after the reproducibility analysis.

The degree of differentiation was also subjected to reproducibility analysis and final scoring in the same way as the classification of the breast cancers. The same procedure was also employed for the WHO grading system.

Results

Table 2 illustrates the inter- and intraobserver reproducibilities of the three classification systems analysed. The interobserver reproducibility varied between 45·5 and 78·6% equal assessments and the intraobserver reproducibility between 57·0 and 75·7%. Tschuprow's test could only be applied in the analysis of Ackerman's system, since in the other two systems the expected number in entries in the contingency table did not exceed 5 in two or fewer numbers—that is, it did not even fulfil the more liberal requirements of Tschuprow's coefficient suggested by Siegel.[16] Cohen's κ varied between 0·24 and 0·49—that is, "fair" or "moderate" according to Landis and Koch.[15]

Table 3 illustrates the inter- and intraobserver reproducibilities of the concepts of high, medium and low degrees of differentiation. This reproducibility is significant but is as low as the reproducibility of the tumour-classification systems. Analogous results were also obtained in the reproducibility analysis of the grading system of WHO. See also Stenkvist *et al.*[12]

Tables 4 and 5 illustrate the lack of correlation between the classification systems of WHO and AFIP and the five-year postmastectomy recurrence rate. In this correlation patients dying of causes other than breast cancer during the follow-up period

Table 3 *Inter- and intraobserver reproducibilities of the degree of differentiation of breast cancer*

	Inter	*Intra 1*	*Intra 2*
Cohen's κ	0·31	0·34	0·45
p value of χ^2 test	0·002	0·003	0·001
Tschuprow's coefficient	0·110	0·196	0·354
% equal assessments	60·8	57·8	71·7
(No of classified tumours)	(166)	(173)	(166)

Table 4 *Correlation between recurrence after five-year follow-up and cancer classification according to Ackerman. Patients dead of causes other than breast cancer and without recurrences within the follow-up period are not included*

Class (Ackerman classification)		*No recurrence*	*Recurrence*
I	Non-invasive	2	0
II 1	Colloid carcinoma	7	2
2	Medullary carcinoma	2	1
3	Well-differentiated adenocarcinoma	5	0
III 1	Adenocarcinoma	27	13
2	Intraductal carcinoma with stromal invasion	50	16
IV	Highly metastasising	14	12

Tschuprow's coefficient = 0·02 with 0·26 associated probability—that is, no correlation between cancer class and recurrence.

Table 5 *Correlation between recurrence rate after five-year follow-up and classification according to Armed Forces Institute of Pathology. Patients dead of causes other than breast cancer without recurrence within the follow-up period are not included*

Class (AFIP classification)		*No recurrence*	*Recurrence*
II b1	Infiltrating papillary carcinoma	5	0
II b2	Infiltrating comedocarcinoma	8	0
II b3	Infiltrating scirrhous carcinoma	72	37
II b4	Infiltrating medullary carcinoma	2	1
II b5	Infiltrating colloid carcinoma	6	2
III b	Infiltrating lobular carcinoma	9	4
Others		6	0

Tschuprow's coefficient = 0·02 with an associated probability of 0·65—that is, no correlation between cancer class and recurrence rate within five years.

Table 6 *Correlation between recurrence rate five-years post-mastectomy and classification according to the World Health Organisation. Patients dead without recurrence of causes other than breast cancer are not included*

Class (WHO classification)		*No recurrence*	*Recurrence*
II	Infiltrating carcinoma	82	37
III a	Medullary carcinoma	2	1
III b	Papillary carcinoma	5	0
III d	Mucous carcinoma	5	2
III e	Lobular carcinoma	9	4
Others		6	0

Tschuprow's coefficient 0·01 with an associated probability of 0·90—that is, no correlation between cancer class and recurrence rate within 5-years post-mastectomy.

were excluded. As a consequence, these tests could be performed on 153 of the patients. Tables 4 and 5 also show that in the classification systems suggested by WHO and AFIP the overwhelming majority of tumours are referred to a single class, leaving only a few breast cancers in the other classes of these comprehensive classification schemes.

Table 6 shows that Tschuprow's test does not discriminate in general between Ackerman's classes. Yet if we compare class IV with the others we get 12/26 versus 32/125 recurrences. On a two-tailed exact test this gives $p = 0.0556$ which suggests a poor prognosis for Ackerman's class IV. However, equally good discrimination is obtained by simply using the concept of differentiation; that is, degree of differentiation was correlated with five-year recurrence rate such that low degrees of differentiation meant a high recurrence rate in Mann-Whitney's U test. The probability of this being a random occurrence was 0·04. Ackerman's class IV was strongly correlated with a low degree of differentiation (28/30 cases).

Similarly, Ackerman's classification correlated well with the WHO gradation system (25·5 in Kruskal-Wallis test with an associated probability of 0·001). Gradation, which subdivides cancer into three groups according to differentiation and aneuploidy, was correlated with recurrence rate such that many points meant a high recurrence rate. The probability of this being a random occurrence was 0·07 in Mann-Whitney's U test. Note that differentiation and gradation do not describe the same biological phenomena.[12]

Discussion

The difficulties in obtaining reproducible, histopathological classifications are well known. The

definition of accuracy of diagnosis, the decision rules, the consistency with which these rules are applied and the factors affecting the consistency have recently been discussed by Langley.[17] The difficulties in arranging quality control procedures in histopathology and cytology are also considered in Langley's study.

No reproducibility analysis or analysis of biological significance has been performed previously for any of the classifications examined here. Excellent reviews of the literature in the field have been given by Schiødt and Fisher[3 8] and by Linell *et al.*[18] Our results do not agree with those of Fisher *et al*, who reported that different intra- and interobserver opinions about diagnoses of the histological type of the tumour occurred in only 3% of 1000 cases in a classification system containing 40 different classes of tumours. These results are better than those reported in our study.

One explanation of our low reproducibility and our difficulty in maintaining the criteria described for different tumour classes may be that we are not sufficiently skilled at classifying breast cancer. However, we believe our results are of average quality as the percent distribution among different tumour classes are in accordance with other studies.[18] In addition, the tumours were diagnosed as a part of a scientific project, and the use of special stains, the definition of the different parameters and the characteristics of the various tumours and systems were all thoroughly discussed before the registration forms were designed. Furthermore, the tumours were reviewed within a short time, a factor that should also contribute to a uniform judgement of the material. Mistakes in notation on the data sheets and errors in transferring the data to the computer were minimised by using built-in controls and by transferring everything twice. Also, a classification system that cannot be applied by anyone but a few highly skilled pathologists is in itself of limited value.

None of the classification systems examined here was found to be correlated with the five-year recurrence rate. The low reproducibility and lack of biological significance, as expressed by the five-year recurrence rate means that they are useless in a clinical situation.

The only tumour class that seemed to have prognostic implications was Ackerman's class IV, but this system was improved simply by dividing primary tumours according to degree of differentiation. The fact that invasive breast carcinomas when subdivided according either to degree of differentiation or gradation according to WHO showed a significant correlation with five-year recurrence indicates that there are characteristics of the tumour itself that are of biological significance. One such characteristic is probably associated with the cell membrane and cytoskeleton (differentiation) and the other associated with the mitotic potential (variability of DNA content in tumour cell nuclei). We have recently shown that these two factors do not describe the same intrinsic tumour properties.[12]

If a system is to be useful in individual prognosis at the time of diagnosis, it must also take into account factors other than the tumour's intrinsic properties—for example, axillary involvement, and tumour size. We have recently described how a combination of factors, all of which are independently important for prognosis, can be combined by step-wise, logistic regression analysis to predict correctly recurrent disease in at least 90% of breast cancer patients on an individual basis.[19] We think that future efforts should be directed towards development of objective measurements of relevant variables in clinical oncology in order to determine prognostic features for clinical use. The prerequisite methods for this are available today.

We acknowledge the support of the National Cancer Institute and the National Institutes of Health (contract No NO1-CB-53968) USA.

References

1 Richardson WW. Medullary carcinoma of the breast. A distinctive tumour type with a relatively good prognosis following radical mastectomy. *Br J Cancer* 1956;**10**:415–23.

2 Silverberg SG, Kay S, Chitale AR, Levitt SH. Colloid carcinoma of the breast. *Am J Clin Pathol* 1971;**55**:355–63.

3 Schiødt T. *Breast carcinoma. A histologic and prognostic study of 650 followed-up cases.* Copenhagen: Munksgaard, 1966.

4 Scarff RW, Torloni H. Histological typing of breast tumours. *International histological classification of tumours. No 2.* Geneva: WHO, 1968.

5 Stewart FW. Tumors of the breast. *Atlas of tumor pathology* sect IX. Washington: Armed Forces Institute of Pathology, 1950.

6 Ackerman LV, Del Regato JA. *Cancer: diagnosis, treatment and prognosis.* 4th ed. Mosby Co, 1970.

7 McDivitt RW, Stewart FW, Berg JW. Tumors of the breast. *Atlas of tumor pathology.* Washington: Armed Forces Institute of Pathology, 1968.

8 Fisher ER, Gregorio R, Fisher B, Redmond C, Vellios F, Sommers C. The pathology of invasive breast cancer. *Cancer* 1975;**36**:1–85.

9 Tobon H, Salazar H. Tubular carcinoma of the breast. Clinical, histological and ultrastructural observations. *Arch Pathol Lab Med* 1977;**101**:310–6.

10 Hultborn KA, Törnberg B. Mammary carcinoma. The biological character of mammary carcinoma studied in 517 cases by new form of malignancy grading. *Acta Radiol* 1960;suppl 196.

11 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.

12 Stenkvist B, Westman-Naeser S, Vegelius J, *et al.* Analysis of the reproducibility of subjective grading systems for breast carcinoma. *J Clin Pathol* 1979;**32**:979–85.

13 Tschuprow AA. *Principles of the mathematical theory of correlation.* London: W Hodge, 1939.

[14] Cohen J. A coefficient of agreement for nominal scales. *Education and Psychol Measurement* 1960;**20**:37–46.
[15] Landis JR, Koch EG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
[16] Siegel S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
[17] Langley IA. Quality control in histopathology and diagnostic cytology. *Histopathology* 1978;**2**:3–18.
[18] Linell F, Ljungberg O, Andersson I. Breast carcinoma. Aspects of early stages, progression and related problems. *Acta Pathol Microbiol Scand* 1980;**suppl 272**:1–156.
[19] Stenkvist B, Bengtsson E, Dahlqvist B, *et al.* Predicting breast-cancer recurrence. *Cancer* 1982;2884–93.

Requests for reprints to: Dr B Stenkvist, Department of Clinical Cytology, University Hospital, S-751 85, Uppsala, Sweden.