2. Sainsbury JRC, Farndon JR, Needham GK, Malcolm AJ, Harris AL. Epidermal growth factor receptor status as predictor of early recurrence of and death from breast cancer. *Lancet* 1987, **i**, 1398–1402.
3. Harris AL, Sainsbury JRC, Smith K, Neal DE, Hall RR, Farnon JR. Epidermal growth factor receptors in primary human breast and bladder cancer: relation to tumor differentiation, invasion and patient survival. In: Klijn JCM, Paridaeus R, Fockens JA, eds. *Hormonal Manipulation of Cancer.* New York, Raven Press, 1987, 415–424.
4. Grimaux M, Romain S, Remwikos Y, Martin PM, Magdelenat H. Prognostic value of epidermal growth factor receptor in node positive breast cancer. *Breast Cancer Res Treat* 1989, **14**, 77–90.
5. Spyratos F, Delarue JC, Andrieu C, *et al.* Epidermal growth factor receptors and prognosis in primary breast cancer. *Breast Cancer Res Treat* 1990, **17**, 83–90.
6. Bolla M, Chedin M, Souvignet C, Marron J, Arnould C, Chambaz E. Estimation of epidermal growth factor receptor in 177 breast cancers—correlation with prognostic factors. *Breast Cancer Res Treat* 1990, **16**, 97–102.
7. Sokal R, Rohlt F. *Biometry.* New York, Freeman, 1981.
8. Delarue JC, Friedman S, Mouriesse H, May-Levin F, Sancho-Garnier H, Contesso G. Epidermal growth factor receptor in human breast cancer: correlation with estrogen and progesterone receptors. *Breast Cancer Res Treat* 1988, **11**, 173–178.
9. Battaglia F, Scambia G, Rossi S, *et al.* Epidermal growth factor receptor in human breast cancer: correlation with steroid hormone receptors and axillary lymph node involvement. *Eur J Cancer Clin Oncol* 1988, **24**, 1685–1690.
10. Cappelletti V, Brivio M, Miodini P, Granata G, Coradini D, Di Fronzo G. Simultaneous estimation of epidermal growth factor receptors and steroid receptors in a series of 136 resectable primary breast tumours. *Tumor Biol* 1988, **9**, 200–211.
11. Rios MA, Macias A, Perez R, Lage A, Skoog L. Receptors for epidermal growth factors and estrogens as predictors of relapse in patients with mammary carcinoma. *Anticancer Res* 1988, **8**, 173–176.
12. Nicholson S, Richard J, Sainsbury C, *et al.* Epidermal growth factor receptor: results of a 6 year follow-up study in operable breast cancer with emphasis on the node negative subgroup. *Br J Cancer* 1991, **63**, 146–150.

Acknowledgements—The authors wish to thank Mrs A. Daru and N. Blanchard for their secretarial help. This work was supported by a grant from Region Rhône-Alpes, Comité Départemental de la Ligue Contre le Cancer, Caisse d'Assurance Maladie des Professions Libérales, Province, France.

Eur J Cancer, Vol. 28A, No. 6/7, pp. 1054–1058, 1992.
Printed in Great Britain

0964–1947/92 $5.00 + 0.00
© 1992 Pergamon Press Ltd

Inter-observer and Intra-observer Variability of Mammogram Interpretation: a Field Study

Giovannino Ciccone, Paolo Vineis, Alfonso Frigerio and Nereo Segnan

To evaluate the performance of radiologists in mammographic mass screening, seven radiologists read blindly the mammograms of 45 women (two views for each breast). The films included 12 normal, 24 benign disease and 9 cancers. The readings were repeated after 2 years. As expected, variability was higher among radiologists than between the two readings of the same radiologist, but general reproducibility was moderate. Kappa values for a positive/negative classification were 0.45 at the first and 0.44 at the second reading (inter-observer comparisons). For the intra-observer comparisons, Kappa values ranged from 0.35 to 0.67 (mean 0.56). Generally, accuracy was low partly due to the difficulty of the cases. A slight increase in sensitivity was observed at the second reading. The level of agreement is a good indicator of accuracy. Proper training and standardization of criteria are essential before mass breast screening is implemented.
Eur J Cancer, Vol. 28A, No. 6/7, pp. 1054–1058, 1992.

INTRODUCTION

It is well accepted that periodic mammographic screening has the potential to reduce mortality rates for breast cancer by a significant amount, at least in women aged 50 or more [1, 2].

In parallel with the implementation of mass breast screening there is increasing interest in improving the validity of the diagnostic procedures involved [2, 3]. Whereas accuracy has been assessed in previous studies evaluating effectiveness of mass breast screening, less is known about inter- and intra-observer reproducibility on mammographic interpretation. In a large bibliography of publications on observer variability [4], only one out of 51 references included in the section on conventional radiology, considered mammography [5].

Where mass screening has not been implemented, knowledge of the actual accuracy and variability among observers in interpreting the screening test could be of interest. In Turin (a northern Italian city with a population of about 1000000) the local health authority is planning a population screening for breast cancer by mammography. We were asked to describe and evaluate activities that might be involved in the screening programme, including the performance of the radiology units

Correspondence to G. Ciccone.
G. Ciccone and P. Vineis are at the Unit of Cancer Epidemiology, Department of Biomedical Sciences and Human Oncology, University of Turin, Via Santena, 7, 10126 Turin; A. Frigerio is at the Radiology Unit, S. Giovanni Hospital, Via Cavour, 31, 10123-Turin; and N. Segnan is at the Epidemiology Unit, Local Health Unit No. 1, National Health Service, Via S. Francesco da Paola, 31, 10123 Turin, Italy.
Revised 4 July 1991; accepted 24 Dec. 1991.

of the city with experience in routine mammography. The present report describes an *ad hoc* study organised to address this problem.

MATERIALS AND METHODS

Radiologists and films

In 1986 an invitation to participate in a study on diagnostic reproducibility and on accuracy assessment in the interpretation of standard mammograms was sent to all public centres in Turin peforming mammography. 8 radiologists agreed to participate, representing five out of the seven centres contacted. Each received the mammograms of 45 women (two views for each breast) without any other information and were asked to classify: (i) the parenchymal pattern in four categories (adipose/ mixed/ dense/not interpretable), (ii) the most probable diagnosis in five categories (normal/diffuse benign disease/nodular benign/suspicion of cancer/cancer) and (iii) the recommendation for further diagnostic procedures in four categories (no recommendation/clinical examination/needle biopsy/surgical biopsy).

The 45 women were selected from subjects screened in the National Breast Screening Study of Canada (NBSS), and their mammograms were kindly made available by Dr C.J. Baines (NCIC-Toronto). The final diagnoses, provided by the same source, were: 9 histologically confirmed cancers, 24 benign diseases, diagnosed after needle biopsy (i.e. false positive cases in the original set), and 12 normal mammograms (on which both the screening centre radiologist and a reference radiologist raised no suspicion of cancer and agreed that the woman need not be seen by a surgeon). Each group of women was a random sample of the corresponding source population.

A preliminary report on the variability among these 8 radiologists has been published [6].

In 1988, about 2 years after the first reading, the same radiologists were asked to participate in the second phase of the study, aiming at evaluating the intra-observer variability. 7 of the 8 radiologists accepted. The same set of 45 mammograms and forms from the previous phase were used. The most important difference in this second reading was that the radiologists were aware of the distribution of diagnoses within the set of films.

Analysis

The Kappa statistic was used to obtain a measure of agreement between each pair of observers or, for the same observer, between two readings [7]. Specific Kappa was computed for each single category (or grouped categories) of diagnosis and of recommendation. The interpretation of the radiologists was considered positive when they either diagnosed or suspected a cancer and whenever they recommended a biopsy. An overall weighted Kappa [8] was computed to have a synthetic measure of general agreement over the entire classification. The weighting procedure allows for the assignment of a 'weight' to take into account the level of disagreement when more than two categories are available for rating. The weights used are those proposed by Cicchetti for dichotomous ordinal scales, i.e. classifications containing an 'absence' point and two or more levels of 'presence' of the rated variable [9]. In each comparison, agreement has been measured only on those mammograms considered interpretable by both observers.

In assessing accuracy we considered false positives, both normal women and those who had a pathological diagnosis of benign disease, as ascertained in the NBSS, positively rated by the radiologists. False negatives were women with a pathological

Table 1. Distribution of the 45 women according to diagnosis and recommendation for further tests, made by each radiologist at the first and the second reading, respectively

	Radiologist (1st and 2nd reading)													
Diagnosis:	A		B		C		D		E		F		G	
- recommendation	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Normal														
No recommendation	9	8	7	6	10	11	4	5	12	9	8	7	7	4
Clinical examination					1				1				1	1
Diffuse benign disease														
No recommendation	7	13	1	1	5	5	3	5	8	2	1	5	5	
Clinical examination			5	4	9	10	12	10	2	5	7	2	7	12
Nodular benign disease														
No recommendation	2	1		1								3		
Clinical examination	4	4	12	10	3	4	5	13	6	3	12	9	3	5
Needle biopsy	6	4	4	3	2	3	5			7		1	1	3
Surgical biopsy			1											
Suspect of cancer														
Clinical examination								1			7	2	1	
Needle biopsy		6	2	7	5	1	2	5		16	1	13	4	4
Surgical biopsy	13	6	6	7	7	9	9	1	10		5	1	5	7
Cancer														
Surgical biopsy	4	3	4	4	3	2	3	4	4	3	3	2	4	2
Technically unsatisfactory			3	2			2	1	2		1		7	7

Table 2. Inter-observer agreement: specific Kappa for agreement on positive/negative categories for each pair of radiologists at the first and the second reading

Radiologist	B		C		D		E		F		G	
	1	2	1	2	1	2	1	2	1	2	1	2
A	0.28	0.55	0.56	0.81	0.58	0.45	0.54	0.52	0.51	0.54	0.53	0.58
B			0.18	0.36	0.33	0.23	0.08	0.32	0.37	0.36	0.18	0.18
C					0.66	0.50	0.59	0.45	0.61	0.53	0.61	0.46
D							0.50	0.23	0.61	0.25	0.51	0.44
E									0.49	0.70	0.31	0.50
F											0.50	0.32

diagnosis of cancer whose mammograms were considered negative by the radiologists.

RESULTS

Variability

Table 1 gives an overview of the results of the dual reading of the mammograms, according to the diagnosis made and the follow-up diagnostic procedure recommended by each radiologist. The number of films considered as technically unsatisfactory was low, except for radiologist G, who systematically excluded all those mammograms showing even slight artefacts on the films. The number of mammograms interpreted as normal ranged from 4 to 13 at the first reading and from 5 to 11 at the second one.

The inter-observer variability for the recommendation of further diagnostic procedures is fairly large: women receiving no recommendations ranged from 7 to 20 at the first reading and from 4 to 22 at the second one. The number of subjects to which either clinical examination, aspiration or surgical biopsy was recommended varied in a similar way. However, if the recommendation of a bioptic test—aspiration or surgical biopsy—is considered as a single category, inter-observer variability is markedly reduced.

There were some systematic differences between the two reading sessions, particularly evident regarding the diagnostic procedure recommended. The average percentage of subjects referred for aspiration was about 10.7% at the first reading and 23.8% at the second. An inverse trend is observed for surgical biopsy, decreasing from 26.9 to 16.9%. These changes between readings are particularly evident for radiologists E and F. Considering aspiration and surgical biopsy together, the difference between readings is reduced from 37.6 to 40.7%.

Table 2 shows the level of agreement on the positive/negative categories (as defined in Materials and Methods) for each pair of radiologists at both readings. The Kappa values ranged from 0.08 (between radiologists B and E, first reading) to 0.81 (radiologists A and C, second reading). Most comparisons showed moderate agreement, with an average Kappa value of 0.45 (S.D. 0.16).

Table 3 reports the means of the Kappas for inter-observer agreement on different categorizations: positive/negative, specific diagnostic procedure recommended and overall agreement on the original four point classification proposed. In general, the level of agreement was moderate and does not vary between readings. Radiologist B has the lowest inter-observer agreement at both readings.

Table 4 reports the Kappas for intra-observer agreement of each radiologist at the repeated reading for the considered categories. Intra-observer is generally greater than inter-observer agreement, with an average increase of Kappa values of about 0.10. Clinical examination is the specific recommendation choice with lowest reproducibility.

Accuracy

Table 5 shows the distribution of the mammograms according to the number of 'positive' interpretations received and the reference diagnosis. In this table, only the 36 women diagnosed by all radiologists in both readings were included. Between the first and the second session, out of 63 readings from 9 normal women, recommendation of biopsy dropped from 14 to 8. Among the 42 readings from 6 cancer patients, the detection of cases increased from 22 to 26.

Table 3. Inter-observer agreement: specific Kappa for agreement on positive/negative categories, for further diagnostic procedures recommended and weighted Kappa† for overall agreement on recommendations (1st and 2nd reading)*

	Specific Kappa									
			Further diagnostic test recommended							
	Positive or negative		No recommendation		Clinical examination		Biopsy (needle or surgical)		Overall weighted Kappa	
Radiologist	1	2	1	2	1	2	1	2	1	2
A	0.50	0.58	0.44	0.36	0.14	0.19	0.46	0.45	0.47	0.47
B	0.24	0.33	0.53	0.46	0.12	0.09	0.18	0.31	0.34	0.41
C	0.54	0.52	0.39	0.43	0.18	0.21	0.50	0.40	0.43	0.49
D	0.53	0.35	0.45	0.33	0.14	0.11	0.46	0.29	0.48	0.33
E	0.42	0.45	0.46	0.49	0.20	0.27	0.41	0.44	0.42	0.47
F	0.52	0.45	0.41	0.41	0.20	0.28	0.35	0.43	0.39	0.45
G	0.44	0.41	0.32	0.37	0.17	0.24	0.41	0.43	0.38	0.42
Mean Kappa values‡	0.45	0.44	0.43	0.41	0.16	0.20	0.40	0.39	0.42	0.43

* Specific Kappa is the mean of the Kappa values of the six comparisons of each radiologist with the others.

† Overall weighted Kappa is the mean of the Kappa values of the six comparisons of each radiologist with the others on the four category classification of the recommendations.

‡ Mean Kappa is based on all 21 possible comparisons of the 7 radiologists.

Table 4. Intra-observer agreement: specific Kappa for agreement on positive/negative categories, for further diagnostic procedures recommended and weighted Kappa for overall agreement on recommendations between first and second reading*

	Specific Kappa				
		Further diagnostic test recommended			
Radiologist	Positive or negative	No recommendation	Clinical examination	Biopsy (needle or surgical)	Overall weighted Kappa
A	0.65	0.55	0.18	0.65	0.61
B	0.35	0.69	0.12	0.36	0.57
C	0.61	0.66	0.31	0.61	0.59
D	0.61	0.64	0.37	0.56	0.64
E	0.39	0.48	0.40	0.41	0.44
F	0.67	0.33	0.18	0.48	0.46
G	0.61	0.40	0.28	0.61	0.56
Mean Kappa values	0.56	0.54	0.26	0.53	0.55

* Overall weighted Kappa is computed for each radiologist on the four category classification of the recommendations.

Table 5. Distribution of the women according to the reference diagnosis (NBSS) and the number of positive evaluations received from the seven radiologists at the 1st and 2nd reading (only 36 subjects diagnosed by all radiologists in both readings)

Reference diagnosis (1st and 2nd reading)	No. of positive evaluations received								
	7	6	5	4	3	2	1	0	Total
Normal									
1	0	1	0	0	0	3	2	3	9
2	0	0	0	1	0	0	4	4	
Benign breast disease									
1	2	4	2	1	1	2	3	6	21
2	2	5	1	3	1	4	1	4	
Cancer									
1	1	1	1	0	0	1	2	0	6
2	1	0	2	0	3	0	0	0	
Total									36

Table 6 reports the proportion of mammograms correctly rated by each radiologist in the two readings, according to the reference diagnosis. The range of cancer detected is from 3/9 (radiologist B, first reading) to 8/9 (radiologists E and F, second reading). Comparing the performance of the radiologists in the two sessions, there was a slight increase in sensitivity, from 61.5 to 69.1, while specificity was roughly constant, around 65.0. Only radiologist G showed a worsening performance at the second reading, both in sensitivity and specificity.

Table 6. Proportions of women correctly rated by each radiologist (sensitivity and specificity) at the 1st and the 2nd reading, according to the reference diagnosis (NBSS)

Radiologist (1st and 2nd reading)	Reference diagnosis			
	Cancer *	(%)	Normal or benign disease †	(%)
A				
1	6/9	(66.7)	19/36	(52.8)
2	6/9	(66.7)	23/36	(63.9)
B				
1	3/9	(33.3)	21/35	(60.0)
2	7/9	(77.8)	20/35	(57.1)
C				
1	6/9	(66.7)	25/36	(69.4)
2	6/9	(66.7)	27/36	(75.0)
D				
1	5/9	(55.6)	20/34	(58.8)
2	4/9	(44.4)	28/35	(80.0)
E				
1	5/9	(55.6)	25/34	(73.5)
2	8/9	(88.9)	18/36	(50.0)
F				
1	6/9	(66.7)	25/35	(71.4)
2	8/9	(88.9)	25/36	(69.4)
G				
1	6/7	(85.7)	22/31	(71.0)
2	3/6	(50.0)	19/33	(57.6)

* Sensitivity, based on a pathological diagnosis.

† Specificity, based on a radiological evaluation (normal) or a pathological diagnosis (benign disease).

DISCUSSION

Effectiveness of screening is influenced by a number of components, the quality of which is important to know before implementation and to monitor after starting. In the case of breast cancer screening, accuracy and reproducibility of mammogram interpretation are central components requiring special attention [2, 3].

In the present study, both accuracy and reproducibility have been measured on a set of 45 mammograms to evaluate the potential performance of a group of 7 clinical radiologists faced with screening-like problems.

In general, moderate agreement is evident for all the categories considered. The level of concordance is obviously higher for the simpler, screening-oriented dicothomous classification, than when specific recommendations are considered. Low agreement is evident for the recommendation of clinical examination and between the two methods recommended for biopsy. This observation is not surprising, because the decision to refer a woman for biopsy and the choice between the available procedures are usually based on both the clinical examination and the locally available diagnostic options. From the first to the second reading there was a general shift from surgical biopsy to aspiration—particularly evident for radiologists E and F—possibly due to the increased availability of low cost instruments for stereotactic needle biopsy.

As expected, agreement within observers is better than between observers and the actual difference is even greater than what appears from a crude comparison of Kappa values. In fact, intra-observer agreement estimates also include an effect of the time interval between the readings. Such a source of error does not affect the inter-observer comparisons.

A low level of specificity was expected, since 24 out of 36 negative cases were already biopsied in the NBSS.

Between the two readings, there was some improvement in the ability to detect cancer cases by most radiologists. This is unlikely to reflect an increased awareness of the approximate frequency of the 'true' diagnoses, since there was no systematic shift of the frequency of each diagnosis between the two readings (Table 1) and the slight increase of sensitivity was associated to a constant specificity (Table 6).

As in most studies on observer variability, generalization of results and comparisons with similar experiences are problematic. While attempts to find a general interpretation for Kappa values have been made [10], it is also clear that estimates of agreement are influenced by the type of classification used, the number of points in the rating scale, the relative frequency of each condition in the set being reviewed and the weights chosen to obtain an overall evaluation of agreement [11, 12].

In the present study sensitivity and specificity were estimated on a low number of films. In addition, the frequency of different conditions and the difficulty of some cases were not representative of that usually seen in a screened population. Thus, the degree of accuracy cannot be interpreted as an estimate of the actual accuracy in a screening context. Conversely, studying variability in a representative sample of screened women would not have been efficient, because of the extremely low prevalence of positive results.

The present study confirms the need for proper training for radiologists who will be involved in a population screening project. The low agreement about what should be recommended is an indication for the need to reach a consensus on a clear operative protocol, according to the outcome of mammogram interpretation. In particular, there is a need to clarify what the role of the physical examination should be in the screening programme, after its recent evaluation as a single screening method [13].

Even if this study, as with the few similar studies published on the same topic [14–16], is affected by a certain degree of arbitrariness and artefact, some of the suggestions coming out may be of general interest. Firstly, the involvement of clinical radiologists in screening programmes, without proper training, could be a cause of less success than expected. Secondly, the implementation of screening programmes should contain quality assurance procedures to monitor important screening components. In the absence of accuracy evaluations, the level of inter- and intra-observer agreement is a useful indicator of quality of the radiologist performance.

1. Day NE, Baines CJ, Chamberlain J, Hakama M, Miller AB, Prorok P. UICC project on screening for cancer. *Int J Cancer* 1986, **38**, 303–308.
2. Day NE, Miller AB, eds. *Screening for Breast Cancer.* Toronto, Hans Huber, 1988.
3. Witcombe JB. A licence for breast cancer? *Br Med J* 1988, **296**, 909–911.
4. Feinstein AR. A bibliography of publications on observer variability. *J Chron Dis 1985*, **38**, 619–632.
5. Clark RL, Copeland MM, Egan RL, *et al.* Reproducibility of the technic of mammography (Egan) for cancer of the breast. *Am J Surg* 1965, **109**, 127–133.
6. Vineis P, Sinistrero G, Temporelli A, *et al.* Inter-observer variability in the interpretation of mammograms. *Tumori* 1988, **74**, 275–279.
7. Fleiss JL. Statistical Methods for Rates and Proportions, 2nd ed. New York, John Wiley & Sons, 1981.
8. Cohen J. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968, **70**, 213–220.
9. Cicchetti DV. Assessing inter-rater reliability for rating scales: resolving some basic issues. *B J Psychiat* 1976, **129**, 452–456.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977, **33**, 159–174.
11. Maclure M, Willett WC. Misinterpretation and misuse of the Kappa statistic. *Am J Epidemiol* 1987, **126**, 161–169.
12. Mazoyer B, Mary JY. Kappa as an index of reproducibility: distribution under the null-hypothesis. *Rev Epidem Sante Publ* 1987, **35**, 474–481.
13. Baines CJ, Miller AB, Bassett AA. Physical examination. Its role as a single screening modality in the Canadian National Breast Screening Study. *Cancer* 1989, **63**, 1816–22.
14. Chamberlain J, Ginks S, Rogers P, *et al.* Validity of clinical examination and mammography as screening for breast cancer. *Lancet* 1975, **II**, 1026–1030.
15. Boyd NF, Wolfson C, Moskowitz M, *et al.* Observer variation in the interpretation of xeromammograms. *JNCI* 1982, **68**, 357–363.
16. Baines JC, McFarlane DV, Wall C. Audit procedures in the National Breast Screening Study—mammography interpretation. *J Can Assoc Radiol* 1987, **37**, 256–260.

Acknowledgements—We acknowledge all the participating radiologists: P. Burke, F. Fasciano, R. Ferraris, M.G. Genovese, E. Luparia, F. Potenzoni, with special thanks to A. Temporelli and G. Sinistrero who organized the trial. Thoughtful advice was provided by Dr P.G. Duca, Dr N. Caporaso and Prof. B. Terracini.