ORIGINAL ARTICLE

Taylor&Francis healthsciences

Estrogen and Progesterone Receptor Assay in Paraffin-Embedded Breast Cancer

Reproducibility of Assessment

Gunilla Chebil, Pär-Ola Bendahl and Mårten Fernö

From the Departments of Pathology (G. Chebil) and Oncology (P.-O. Bendahl, M. Fernö), University Hospital, Lund, Sweden and the Department of Pathology (G. Chebil), Helsingborg Hospital AB, Helsingborg, Sweden, for the South Sweden Breast Cancer Group and the North Sweden Breast Cancer Group[1]

Correspondence to: Gunilla Chebil, Department of Pathology, Helsingborg Hospital AB, SE-251 87 Helsingborg, Sweden. Tel: +46 42 10 1990. Fax: +46 42 10 1999. E-mail: gunilla.chebil@helsingborgslasarett.se

Acta Oncologica Vol. 42, No. 1, pp. 43–47, 2003

Immunohistochemical (IHC) assays are routinely used for determining the estrogen receptor (ER) and progesterone receptor (PgR) status of women with breast cancer. The present knowledge of reproducibility of the subjective IHC assessment is limited. The purpose of this study was to evaluate the interobserver reproducibility in IHC assessment of 127 (ER) and 126 (PgR) cases of primary breast cancer. The slides were independently re-assessed by 22 pathologists from nine hospitals. A simple assessment protocol was used with a cut-off value $>10\%$ stained nuclei, regardless of staining intensity. The concordance was high, with a kappa value of 0.78 for ER and 0.72 for PgR. Since some of the 22 pathologists considered one or several sections to be inevaluable, the evaluation comprised three groups ($\leq 10\%$ vs. $>10\%$ vs. not evaluated). Although the percentage of positive cases varied among the 22 pathologists (ER: 57%–79%; PgR: 41%–66%), the overall concordance was good for both ER and PgR (kappa = 0.78 and 0.72, respectively). A further improvement in concordance should be possible to achieve through training.

Received 8 April 2002
Accepted 21 November 2002

Estrogen (ER) and progesterone receptors (PgR) are well-accepted predictors for the effect of endocrine therapy in both primary and metastatic breast cancer (1,2). In recent years, the immunohistochemical (IHC) receptor assay has increasingly replaced the biochemical assay. Comparisons between ER status obtained by the biochemical and the IHC assays have shown 81%–96% concordance (3–5), while for PgR status the reported concordance has been 72%–83% (4,6,7). The assessment of staining has been under debate and several semi-quantitative scoring systems have been proposed: the 'Immunocytochemical score' (8), the 'HSCORE' (9), the 'Simplified histoscore' (10), the 'Quick-score' (11) and the 'Working protocol scoring system' (12) are examples of systems adding or multiplying scores for percentage of positive nuclei with staining intensity. There are, however, studies indicating that evaluation of staining intensity does not yield any additional information to the percentage of stained nuclei and that different scoring methods all give similar prognostic information (3,13,14). Regardless of which scoring system is used, it has to be evaluated for assessment reproducibility and subject to quality assurance. So far, investigations concerning the reproducibility of the assessment of staining have been given less attention than studies on the quality of the staining methods. Published investigations include either a limited number of assessors or a small number of cases: 2/102 (15), 13/2 (16), 8/13 (17). Interassessor variability has been reported as high when combining staining intensity and percentage positivity (17).

The aim of the present study was to more extensively investigate the reproducibility of the assessment of both ER and PgR immunostaining in archival routinely treated clinical material of paraffin-embedded breast cancer in a large study including many samples (n = 134) and many assessors (n = 22).

[1] Participating Departments of Pathology and Pathologists: Halmstad (Kristina Regnström, Görel Östberg), Helsingborg (Ole Ankjaergaard, Poul Boiesen, Birgitta Jakobsson), Karlskrona (Erik Holm, Marcella Mandaus, Karin Schemschat), Kristianstad (Ingmar Glifberg), Lund (Henryk Domanski, Ingrid Idvall, Anna Mas-Morillas), Malmö (Lola Anagnostaki, Otto Ljungberg, Janne Malina), Sundsvall (Antti Sahipakka), Umeå (Bela Bozoky, Thomas Höckenström, Göran Landberg), Växjö (Sven Johansson, Ingrid Moqvist, Per-Göran Persson).

 ISSN 0284-186X

MATERIAL AND METHODS

Study design

Nine departments of pathology participated (seven from the South Sweden Health Care Region and two from the North Sweden Health Care Region), representing three university hospitals and six county hospitals. The number of participating pathologists from each of the nine departments varied between 1 and 3 (in total 22 pathologists), most of them being highly experienced, but not specifically in breast pathology or immunohistochemistry. Two of the pathologists were at the time residents. Each department of pathology contributed with 15 consecutive routine archival cases of breast cancer, which had been investigated with immunohistochemical assay for steroid hormone receptors during the first trimester of 1999 (one department included only 14 cases). A total of 134 cases were thus included in the study.

All departments participated in an external quality assessment program for immunohistochemistry and their technical performance of the assay has been assessed as 'good' or 'acceptable'.

Three slides from each case (routine hematoxylin-eosin, ER and PgR) were sent to the coordinator of the study (MF) for coding and distribution. Each pathologist evaluated 15 (or 14) cases on 9 different occasions. After each of the 198 evaluations (22 pathologists × 9 occasions), the 15 (14) cases were sent to the coordinator for re-coding. The whole evaluation procedure lasted for 8 months. Thereafter, a consensus group consisting of five persons (three pathologists and two technicians), all with considerable experience of IHC, assessed all the slides at a multiheaded microscope without previous knowledge of the results of the other 22 pathologists. The entire slides were assessed, except in cases with variations in receptor staining owing either to tumor heterogeneity or to staining properties where hot-spots were assessed (18). The consensus assessment is used hereafter as the reference result and, furthermore, it was used for exclusion of cases not representative of breast cancer, see below. No one in the consensus group was among the 22 pathologists participating in the study.

Analytical methods

The participating laboratories used their routine protocols for both hematoxylin-eosin and immunostaining. Paraffin sections of 4–5 μm were cut, the tissue sections were placed on precoated slides, dried in 58–60° for 60 min, cooled, and deparaffinized. Heat-induced antigen retrieval was performed in 0.01 M citrate buffer pH 6.0 or 7.0–7.3 in all the laboratories with microwave treatment varying from 270W for 3 × 15 min to 750W for 2 × 5 min. Two antibodies were used for the ER assays, clone 1D5, diluted 1:100, (DAKOPATTS AB, Sweden) or clone 6F11, diluted 1:20 (Novocastra Laboratories Ltd, UK) or prediluted (Ventana Medical System, Strasbourg). For the PgR assays, A0098, diluted 1:50 (polyclonal antibody, DAKO) or clone 1A6, diluted 1:100 (Novocastra) or prediluted (Ventana) were used. All assays were performed in immunostainers; four of the laboratories used Ventana ES or NeXes, three laboratories used DAKO Horizon or TechMate, and two used Shandon Cadenza.

The primary aim of this study was to investigate the reproducibility of receptor status (negative vs. positive) In agreement with several other groups we used a cut-off at $>10\%$ stained nuclei in the invasive component of the tumor without consideration of intensity (12,19–21). The protocol included the option of free comments. Since some of the 22 pathologists considered one or several sections inevaluable (no result was reported and the free comment 'not evaluable' was used), a further group 'not evaluable' was introduced. Thus, evaluation of the reproducibility is in the following based on these three groups: Negative ($\leq 10\%$ stained nuclei) vs. positive ($>10\%$ stained nuclei) vs. not evaluable.

The consensus group excluded seven cases for the following reasons: lymphoma (n = 1), only DCIS (n = 4), sclerosing adenosis (n = 1), not a representative section for IHC (n = 1). The statistical analyses were, thus performed in 127 cases. One additional slide was excluded for PgR– assessment because of the poor quality of the tissue section.

Statistics

The degree of concordance between observers classifying a set of samples as negative, positive, or not evaluable was quantified in two ways in this study: as the percentage of concordant samples and as the chance corrected measure of agreement, known as kappa (22). The latter measure is zero if the concordance is equal to that expected by chance, negative if worse and, positive if better. No absolute definitions for the interpretation of different kappa values are possible, but the following guidelines might be useful: <0.20 Poor; 0.21–0.40 Fair; 0.41–0.60 Moderate; 0.61–0.80 Good; and 0.81–1.00 Very good. The mean kappa per assessor was calculated as an average of 21 pair-wise kappa values (number of assessors–1) and the overall mean kappa value as an average of all 231 pair-wise kappa values. When calculating kappa, all possible deviations from perfect concordance were given equal weight.

RESULTS

Reproducibility of assessment of ER

The consensus group assessed 72% (92/127) of the tumors as positive, 24% (31/127) as negative, and 3% (4/127) as not evaluable. Among the 22 participating pathologists, the percentage of ER-positive cases varied between 57% and 79%, the ER-negative cases between 21% and 42%, and the not-evaluable cases between 0% and 6%. All 22 pathologists

Table 1

The best and worst agreement between two pathologists when evaluating ER staining [not evaluated (N.E.) vs. ER negative (ER−) vs. ER positive (ER+)]

	Best agreement			Worst agreement		
	N.E.	ER−	ER+	N.E.	ER−	ER+
N.E.	1	0	0	0	0	0
ER−	1	41	1	0	27	0
ER+	0	0	83	0	26	73

Table 2

The mean (range) pair-wise concordance (% and kappa) between 22 pathologists in evaluating ER and PgR immunohistochemical staining

Group	ER		PgR	
	Concordance %	Kappa value	Concordance %	Kappa value
All pathologists	90	0.78	86	0.72
'Best' pathologist	93 (83–98)	0.83 (0.66–0.94)	89 (77–94)	0.78 (0.65–0.89)
'Worst' pathologist	85 (79–81)	0.68 (0.57–0.78)	80 (75–86)	0.60 (0.52–0.72)

and the consensus group reported concordant results in 61% of the cases (60 positive and 18 negative). Of 49 discordant cases, one or several pathologists assessed 11 as not evaluable. With the results of the consensus group as the reference, the 'best' pathologist reported concordant results in 121 of the 127 cases (95%), whereas the pathologist with the lowest agreement in relation to the consensus group reported only 104 concordant results (82%). Overall, 15 of the 22 pathologists reported ER status concordant with that of the consensus group in more than 90% of the cases. The best and worst pair-wise agreement between two pathologists was 98% and 79%, respectively (Tables 1, 2). The discordant cases for the two pathologists with the lowest agreement were all consistently assessed as negative by one of them and positive by the other. The overall mean kappa value for all 22 pathologists was 0.78, indicating a good reproducibility (Table 2). The pathologist with the best overall concordance with the others had a mean kappa value of 0.83. The corresponding value for the pathologist with the worst concordance was 0.68.

Reproducibility of assessment of PgR

The consensus group assessed 56% (71/126) of the tumors as positive, 39% (49/126) as negative and 5% (6/126) as not evaluable. Among the 22 participating pathologists, the percentage of PgR-positive cases varied between 41% and 66%, the PgR-negative cases between 34% and 57% and the not evaluable between 0% and 7%. All 22 pathologists and the consensus group reported concordant results in 49% of the cases (41 positive and 21 negative). With the results of the consensus group as the reference, the 'best' pathologist reported concordant results in 116 of the 126 cases (92%). The pathologist with the lowest agreement in relation to the consensus group reported only 100 concordant results (79%). Seven pathologists reported PgR status concordant with the consensus group in more than 90% of the cases. The best and worst pair-wise agreement between two pathologists was 94% and 74%, respectively (Tables 2, 3). The same pathologists as for ER reported the lowest and highest percentage of positive cases and the highest number of not evaluated ones. The discordant cases ($n = 31$) for the two pathologists with the worst agreement were, with one exception, all considered to be negative by one of them and positive by the other (Table 3). The overall mean kappa value for all pathologists was 0.72, indicating good reproducibility (Table 2). The pathologist with the highest and lowest overall concordance with the others had mean kappa values of 0.78 and 0.60, respectively.

DISCUSSION

Using a simple assessment protocol with a cut-off value at $>10\%$ stained nuclei, not considering staining intensity, the reproducibility of the assessment of staining was good for both ER and PgR (kappa 0.78 and 0.72, respectively) when 22 pathologists from 9 different departments evaluated the same slides independently. Considering the evaluation of a consensus group as the reference, 15 of the pathologists reported concordant ER results (negative vs. positive vs.

Table 3

The best and worst agreement between two pathologists when evaluating PgR staining [not evaluated (N.E.) vs. PgR negative (PgR−) vs. PgR positive (PgR+)]

	Best agreement			Worst agreement		
	N.E.	PgR−	PgR+	N.E.	PgR−	PgR+
N.E.	1	0	0	0	0	0
PgR−	0	60	2	0	42	1
PgR+	1	4	58	0	30	51

not evaluable) in more than 90% of the cases. Overall, the reproducibility of PgR was lower than that for ER, kappa value 0.72 vs. 0.78. Only seven pathologists reported results concordant with the consensus group in more than 90% of the cases. The number of free comments in the assessment protocol was also higher for PgR (3.8/case) than for ER (1.5/case). The reason for this difference may be a better standardized protocol for ER staining including more reliable antibodies than for PgR. All the laboratories that participated in the study have today abandoned the polyclonal PgR antibody in favor of a monoclonal antibody (clone 1A6 provided by Novocastra and DAKO or clone 636 from DAKO) with higher specificity.

Most of the pathologists with many diverging results consequently reported more negative or more positive cases than the consensus group. This could be corrected through training and frequent discussions at multiheaded microscopes. In clinical practice, consensus assessments should be considered for borderline staining, and for cases that are difficult to interpret. The prerequisite for the present reproducibility study was that the technical performance of the IHC method, including antigen retrieval and staining, was comparable between the participating departments. The variation in the number of cases per department that were not evaluated may conflict with this technical agreement. When we exclude the cases that were assessed as not evaluable by one or more of the pathologists, the reproducibility of the assessment was very good for ER and good for PgR (kappa 0.82 and 0.76, respectively).

The participation in external quality control programs and continuing validation of standardized procedures as well as validation of assessment schemes and assessors should be obligatory for departments routinely assessing ER and PgR for the clinical management of patients with breast cancer. Quality assurance procedures and programs are established for the staining technique, i.e. the UK NEQAS program. Programs for quality assurance of assessment and scoring have still to be established.

ACKNOWLEDGEMENTS

We express our gratitude to all those who together with Gunilla Chebil participated in the consensus group (Christina Andersson, Monica Haglund, Leif Johansson and Bela Veress). The study was supported by the Thelma Zoégas Foundation, Kristianstad University Clinical Research Council, the Hospital of Lund Foundation and the Swedish Cancer Society.

REFERENCES

1. McGuire W. An update on estrogen and progesterone receptors in prognosis for primary and advanced breast cancer. In: Iacobelli S, et al, eds. Hormones and cancer. New York: Raven Press, 1980: 337–43.
2. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. Lancet 1998; 351: 1451–67.
3. Barnes DM, Harris WH, Smith P, Millis RR, Rubens RD. Immunohistochemical determination of oestrogen receptor: comparison of different methods of assessment of staining and correlation with clinical outcome of breast cancer patients. Br J Cancer 1996; 74: 1445–51.
4. Zafrani B, Aubriot M-H, Mouret E, et al. High sensitivity and specificity of immunohistochemistry for the detection of hormone receptors in breast carcinoma: comparison with biochemical determination in a prospective study of 793 cases. Histopathology 2000; 37: 536–45.
5. Allred C, Bustamente M, Daniel C, et al. Immunocytochemical analysis of estrogen receptors in human breast carcinomas. Arch Surg 1990; 125: 107–13.
6. Elledge R, Green S, Pugh R, et al. Estrogen receptor (ER) and progesterone receptor (PgR) by ligand-binding assay compared with ER, PgR and pS2, by immunohistochemistry in predicting response to tamoxifen in metastatic breast cancer: a southwest oncology group study. Int J Cancer (Pred Oncol) 2000; 89: 111–7.
7. Stierer M, Rosen H, Weber R, et al. Comparison of immunohistochemical and biochemical measurement of steroid receptors in primary breast cancer: evaluation of discordant findings. Breast Cancer Res Treat 1998; 50: 125–34.
8. Heubner A, Beck T, Grill H-J, Pollow K. Comparison of Immunocytochemical estrogen receptor assay, estrogen receptor enzyme immunoassay, and radioligand-labeled estrogen. Receptor assay in human breast cancer and uterine tissue. Cancer Res 1986; 46(Suppl): 4291–5.
9. Kinsel LB, Szabo E, Greene GL, Konrath J, Leight GS, McCarty Jr KS. Immunocytochemical analysis of estrogen receptors as a predictor of prognosis in breast cancer patients: comparison with quantitative biochemical methods. Cancer Res 1989; 49: 1052–6.
10. Benítez-Bribiesca L, Guevara R, Riuz MT, Martínez G, Rodríguez-Cuevas S. A simplified histoscore for the estrogen

receptor assay in breast cancer. Path Res Pract 1992; 188: 461–5.
11. Detre S, Saccani Jotti G, Dowsett M. A 'quickscore' method for immunohistochemical semiquantitation: validation for oestrogen receptor in breast carcinomas. J Clin Pathol 1995; 48: 876–8.
12. Leake R, Barnes D, Pinder S, et al. Immunohistochemical detection of steroid receptors in breast cancer: a working protocol. J Clin Pathol 2000; 53: 634–5.
13. Fernö M, Andersson C, Fallenius G, Idvall I. Oestrogen receptor analysis of paraffin sections and cytosol samples of primary breast cancer in relation to outcome after adjuvant tamoxifen treatment. Acta Oncol 1996; 35: 17–22.
14. Rhodes A, Jasani B, Balaton AJ, Barnes DM, Miller KD. Frequency of oestrogen and progesterone receptor positivity by immunohistochemical analysis in 7016 breast carcinomas: correlation with patients' age, assay sensitivity, threshold value, and mammographic screening. J Clin Path 2000; 53: 688–96.
15. Biesterfeld S, Veuskens U, Schmitz F-J, Amo-Takyi B, Böcking A. Interobserver reproducibility of immunocytochemical estrogen- and progesterone receptor status assessment in breast cancer. Anticancer Res 1996; 16: 2497–500.
16. Van Diest PJ, Weger DR, Lindholm J. Reproducibility of subjective immunoscoring of steroid receptor in breast cancer. Anal Quant Cytol Histol 1996; 18: 351–4.
17. McClelland RA, Wilson D, Leake R, Finlay P, Nocholson RI. A multicentre study into the reliability of steroid receptor immunohistochemical assay quantification. Eur J Cancer 1991; 27: 711–5.
18. Biesterfeld S, Klüppel D, Koch R, et al. Rapid and prognostically valid quantification of immunohistochemical reactions by immunohistometry of the most positive tumour focus. A prospective follow-up study on breast cancer using antibodies against MIB-1, PCNA, ER and PR. J Pathol 1998; 185: 25–31.
19. Goldhirsch A, Glick JH, Gelber RD, Senn HJ. Commentary: International Consensus Panel on the Treatment of Primary Breast Cancer. J Nat Cancer Inst 1998; 90: 1601–8.
20. Molino A, Micciolo R, Turazza M, et al. Prognostic significance of estrogen receptors in 405 primary breast cancers: a comparison of immunohistochemical and biochemical methods. Breast Cancer Res Treat 1997; 45: 241–9.
21. Pertschuk L, Feldman J, Yong-Doo K, et al. Estrogen receptor immunocytochemistry in paraffin embedded tissues with ER1D5 predicts breast cancer endocrine response more accurately than H222Spγ in frozen sections or cytosol-based ligand-binding assays. Cancer 1996; 77: 2514–9.
22. Altman DG. Practical statistics for medical research. London, UK: Chapman & Hall, 1991: 403–9.