Review Article · Übersichtsarbeit

Breast Care

Breast Care 2008;3:128–133
DOI: 10.1159/000121885

Published online: April 15, 2008

Implementation of External Quality Assurance Trials for Immunohistochemically Determined Breast Cancer Biomarkers in Germany

Reinhard von Wasielewski[a] Claudia A. Krusche[a] Joseph Rüschoff[b] Anette Fisseler-Eckhoff[c] Hans Kreipe[a]

[a] Institut für Pathologie, Medizinische Hochschule Hannover,
[b] Institut für Pathologie, Städtisches Klinikum Kassel,
[c] Institut für Pathologie, Horst-Schmidt Kliniken, Wiesbaden, Germany

Key Words
Steroid receptor · HER2 · Immunohistochemistry · Quality assurance

Summary
Besides typing and grading of breast cancer, pathologists are involved in the determination of biomarkers, such as steroid hormone receptors and HER2, which are of utmost importance in adjuvant therapy. There have been concerns with regard to security and reproducibility of the biomarker assays done on tissue sections applying either immunohistochemistry or in-situ hybridisation. In order to assure the quality of these biomarker assays, a number of measures are required, among them external proficiency testing. Therefore, external quality assurance trials have been implemented in Germany. In the period of 2002–2007, 5 consecutive trials were conducted with up to 180 participating laboratories. Tissue microarrays with 20–24 different breast cancer samples including cell lines enabled that a huge number of pathologists were challenged with identical samples which provides the prerequisite for comparability. Because there is no legal duress to undergo external proficiency testing in histopathology, all laboratories that took part volunteered to do so. These innovative quality assurance trials (Qualitätsinitiative Pathologie, QuIP) will be continued in the future on an annual or bi-annual basis. Participation is recommended for pathology departments involved in the service for breast units. The organisational frame work of the trials is described here.

Schlüsselwörter
Steroidhormonrezeptoren · HER2 · Immunhistochemie · Qualitätssicherung

Zusammenfassung
Über die histologische Typisierung und Graduierung hinaus, hat die Pathologie in der Brustkrebsdiagnostik die Aufgabe, Zielstrukturen, die in der adjuvanten Therapie von großer Bedeutung sind, wie Steroidhormonrezeptoren und HER2, zu bestimmen. Die Reproduzierbarkeit dieser quantitativ bzw. semiquantitativ durch Immunhistochemie oder In-situ-Hybridisierung ermittelten Parameter wird zunehmend als eine Aufgabe der Qualitätssicherung wahrgenommen. Um die Qualität dieser Biomarker-Assays zu gewährleisten, kommen verschiedene Maßnahmen in Betracht, unter anderem die Teilnahme an externen Ringversuchen. In Deutschland wurden derartige Ringversuche ins Leben gerufen und fanden in den Jahren 2002–2007 bereits fünfmal mit bis zu 180 teilnehmenden Laboren statt. Dabei kommen «Gewebe-Arrays» (Tissue microarrays) mit 20–24 Gewebeproben von verschiedenen Mammakarzinomen und auch Zelllinien zum Einsatz. Hiermit gelingt es, ein hinsichtlich der Zusammensetzung vielfältiges, hinsichtlich der Anforderungen an die einzelnen Teilnehmer jedoch nahezu identisches Testmaterial an eine hohe Teilnehmerzahl zu distribuieren, wodurch Vergleichbarkeit hergestellt werden kann. Da für eine Ringversuchsteilnahme in der diagnostischen Pathologie keine gesetzlichen Verpflichtungen bestehen, erfolgt die Teilnahme freiwillig. Diese innovative Form von Ringversuchen (Qualitätsinitiative Pathologie, QuIP) wird auch in Zukunft mit einem ein- oder halbjährigen Turnus fortgesetzt werden. Die Teilnahme ist für zertifizierte Brustzentren empfohlen. Hier werden Voraussetzungen und der organisatorische Rahmen dargestellt.

KARGER

Fax +49 761 4 52 07 14
E-mail Information@Karger.de
www.karger.com

Accessible online at:
www.karger.com/brc

Prof. Dr. med. Hans H. Kreipe
Institut für Pathologie
Medizinische Hochschule Hannover
Carl Neuberg Strasse 1, 30625 Hannover, Germany
Tel. +49 511532-4450, Fax -5799
E-mail Kreipe.Hans@mh-hannover.de

Introduction

For decades, clinical cancer research focussed on the study of empirical combinations of non-specific cytotoxic drugs. In recent years, oncology is witnessing a revolution sparked by targeted therapies, notably the chimeric monoclonal antibodies against surface molecules such as CD20 or epidermal growth factor receptor. Today, almost all patients suffering from B-cell lymphomas are treated with this mode of therapy [1]. How does this revolution of therapy interfere with the classical function of histopathology to classify and to grade malignant neoplasms? Will morphological categories be replaced by a list or profile of markers which constitute potential targets for therapy? This will certainly not be the case, although the biological significance of lymphoma classification has to be reconsidered against the background of treatment response which will potentially be more relevant than the spontaneous course of disease. Whereas the task of typing and grading will still form the indispensable basis of cancer therapy, new additional challenges with regard to reliability and reproducibility of target identification are awaiting modern pathology. In particular, quantitative parameters might be insufficiently reproducible.

Breast cancer in recent years has functioned as pioneer tumour, setting the stage for a new era of diagnostic and therapy in oncology. Steroid hormone receptors and the human epidermal growth factor receptor 2 (HER2) provided the first examples for targeted therapy, and marked the beginning of the age of personalised medicine. There are different modes of determining potential target molecules in cancers. Besides tissue extract-based quantitative protein and mRNA assays, there are in-situ methods which apply immunohistochemistry (IHC) or different modes of in-situ hybridisation (fluorescent (FISH), chromogenic (CISH), silver-enhanced (SISH)). In most countries, the latter methods are predominantly used to assess target molecules in breast cancer. However, there are a number of caveats and open issues which have to be kept in mind when in-situ techniques are applied. First, the demand for quantification has to be met, and thresholds for categorisation have to be defined [2]. The biological significance and justification of these thresholds is particularly unclear in the grey zone between unequivocal positive and negative cases [3–5]. Second, the issue of reproducibility and reliability of these assays emerges. The findings of a number of studies indicate that significant interlaboratory variability for steroid hormone receptor and HER2 testing does occur [5–7]. Despite these potential hazards, IHC offers a number of decisive advantages like correlation to number of tumour cells and their viability as well as to admixture of normal, non-invasive, and stromal cells. In addition, alternative extract-based methods did not yet prove a higher degree of reproducibility when applied on a similarly large scale like IHC. Apart from these considerations, pathologists who apply IHC and clinicians who rely on the results of IHC assays need information on how secure – with regard to sensitivity and specificity – the method in individual use really is.

Biomarkers in Breast Cancer

Steroid hormone receptor expression is one of the most important biomarkers in breast cancer, which provides the basis for the selection of alternative therapeutic strategies in adjuvant breast cancer treatment [8]. In recent years, HER2 has gained similar impact as prognostic and predictive marker which is meanwhile evaluated on a regular basis and influences therapeutic decisions in the management of breast cancer patients [3]. For several reasons, both biomarkers usually are determined by pathologists applying tissue sections and IHC. In particular, differentiation of invasive cancer cells in heterogeneous tissue encompassing normal epithelial cells, stroma and potentially in situ lesions or necrosis requires microscopic correlation. Immunohistochemical biomarker assays, however, do not represent a simple extension of traditional histopathological evaluation because it includes quantitative assays whereas the traditional and unquestioned strength of histopathology lies in qualitative analysis. In order to cope with the new challenge of target molecule detection in the age of personalised medicine, pathologists have to prove that quantitative biomarker assays done by them on breast cancer tissue are accurate and reliable.

Testing inaccuracy remains a major issue with both IHC and FISH, and it has been estimated that approximately 20% of current HER2 testing may be incorrect [3]. There is widespread concern that inaccuracy of detection methods and interpretation may lead to an unacceptably high error rate in determining the true hormone receptor status [4]. Comparison of centrally versus locally assessed oestrogen receptor (ER) and progesterone receptor (PR) revealed divergent results in a substantial proportion of patients [5]. Obviously, there is a great need to standardise immunohistochemical biomarker assays to further ensure that similar results are obtained by different institutions.

Tissue Microarrays for External Proficiency Testing

Principally, there are 2 ways to cope with the problem of inaccuracy and poor reproducibility: centralisation of diagnostics or standardisation of diagnostics in a multicentre setting. In Germany, pathologists have decided to opt for the second alternative, and consequently nation-wide trials for tissue-based markers in breast cancer have been set up [9, 10].

External proficiency testing has been proposed as one potential instrument to enable accurate biomarker determination in a non-centralised approach [3, 11, 12]. Yet, the most effective setting for external proficiency testing has not been deter-

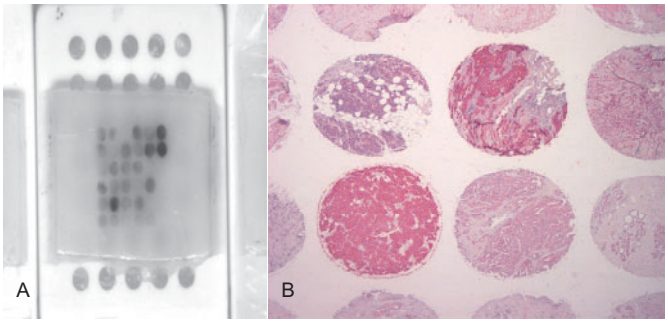

Fig. 1. **A** Paraffin block of a tissue array which is used in the immunohistochemical quality assurance trial; **B** 30 different tumour samples with defined target expression are assembled in one slide which has been stained for cytokeratin. Up to 200 slides can be produced from one tissue array assuring that all participants in the trial obtain almost identical material and that results among different laboratories become comparable. In the quality network of the German Society for Pathology and the Berufsverband Deutscher Pathologen ('QuIP', *www.ringversuch.de*; *www99.mh-hannover.de/institute/pathologie/dgp*), quality assurance trials based on tissue arrays have been set up for different target molecules (ER, PR, HER2, c-kit).

mined. Open issues refer to selection of material to be distributed, adequate number of challenges (cases), type of challenge (cell lines, cancer tissue), and mode of evaluation. During the years of 2002–2007, 5 tissue microarrays (TMA, fig. 1) were generated and distributed to participating laboratories in Germany on demand. Tissue cores from routine surgical pathology samples retrieved from the archives of 3 Institutes of Pathology in Germany (Hanover, Kassel, Wiesbaden) were used for the construction of TMA. Cases were retrieved from the archives with particular emphasis on low steroid hormone receptor expression (Allred score 3–4) [13] and borderline positivity for HER2 (2+). Besides equivocal cases, clearly positive or negative samples were included. Only samples that received identical testing in all 3 laboratories mentioned above entered the final trial. The procedure of selecting the adequate material is depicted in figure 2. Between 20 and 24 samples were included in the TMAs which were generated exactly as described [14]. Pathology departments volunteering to participate in external proficiency testing could order up to 3 test' slides which were freshly cut and shipped unstained. Within 2 months, immunohistochemical stains had to be performed and a protocol of the assessment as well as the stained slides had to be returned to the organisers of the trial. Participants were free to perform only 1 of the 3 tests, or all of them. Unstained slides could be ordered during a 10-month period during which the trial was open for participation. In the 2006 run, for example, the slides contained tissue spots with high expression (ER, PR: n = 6; HER2 3+: n = 7), medium (ER, PR: n = 6; HER2 2+: n = 7), low (ER, PR: n = 6; HER2 1+: n = 4), or no expression (ER, PR: n = 5; HER2 0: n = 4).

With the help of tissue arrays, it becomes possible for the first time to distribute several tumours among a high number of participating pathologists whereby almost identical tumour areas will be studied by all participants. The first and the final slide sectioned from a tissue array block have a distance of less than 1 mm. Furthermore, potential hazards by tumour heterogeneity are neutralised by a high number of samples which are encompassed by a tissue array.

Besides the German Qualitätsinitiative Pathologie (QuIP), other systems for interlaboratory proficiency testing are available in Europe. These systems do not rely on TMA, and only a limited number of samples can be distributed. In the UK, NEQAS-ICC has been founded [11] with only a very small number of German laboratories participating. In Scandinavia, NordiQC has been established, which is similar to the UK NEQAS.

Interobserver and Interlaboratory Variability

Unlike proficiency testing in clinical chemistry, there are 2 potential sources of error in immunohistochemical assays. On the one hand, the staining may be insufficient due to shortcomings in the immunohistochemical laboratory. On the other hand, a correct stain may be inadequately evaluated by an inexperienced pathologist. Therefore, both aspects have to be regarded. This requires microscopic re-evaluation of every staining performed by the participants. Consequently, participants do not only fill in a formula with their results but return the slides which were processed in their laboratory. Although this procedure of central review is tedious, it makes sure that the effective performance of an individual immunohistochemical laboratory is controlled. In fact, in most trials, rather the laboratory performance than the microscopic evaluation by the participating pathology department was responsible for high or low levels of concordance. Poor interlaboratory agreement usually is based on insufficient retrieval efficacy or suboptimal IHC.

Aberrations from an expected result can be differently severe ranging from light deviation to completely wrong. Therefore, the central re-evaluation applies a grading scheme to assess deviations. The grading scheme consisted of a 4-tiered score which is applied to every tissue spot in the TMA. Complete accordance with the expected result is scored as 3 points, mild deviation (e.g. Allred score 6 instead of 8) as 2 points, more severe deviations (e.g. Allred score 3 instead of 8) as 1 point, whereas false-positive or -negative results are always scored as 0 (fig. 3). In a tissue array with 24 samples, the maximum sum of score points is 72. Percentage values of the maximum score are reported to participating laboratories and provide the basis for benchmarking. Participants fill out an accompanying questionnaire in order to gather information about antigen retrieval and detection methods. The correlation of methods applied and performance in the trial is communicated to all participants in order to enable improvement in those institutions which scored below average.

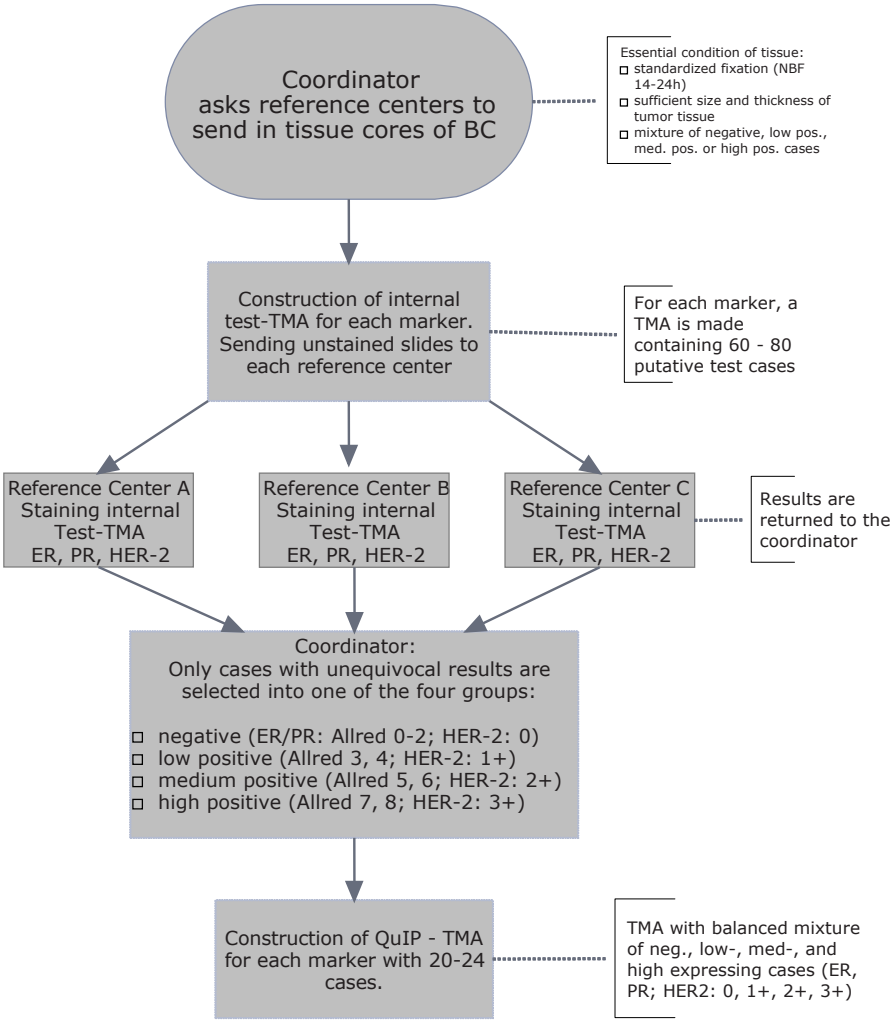

Fig. 2. Case selection for QuIP test. During a pre-test, adequate material is selected to be used for the trial. Only tissues which receive unequivocal and identical testing in all 3 reference laboratories will enter the final tissue microarray (TMA) which will be distributed to the participants.

Benchmarks and Selection of Challenges

Once the concordance rate has been determined by central review, the challenge emerges to set benchmarks discriminating failure from success. In the literature, different thresholds ranging from 80 to 90% are reported [3, 15]. However, these thresholds have to be considered as arbitrary, and currently there is no sufficient data base to define what is sufficient. In addition, there is no doubt that the composition of challenges/cases is of pivotal importance for the outcome with regard to proportion of underscoring laboratories. The more borderline cases and weakly positive case are included in the trial, the lower the concordance rates that can be expected [16]. Accordingly, every TMA should have its own benchmark, which is however impracticable. Alternatively, as has been described by Fitzgibbons et al. [15], benchmarks could be obtained by evaluating the results of all participants which then provide the basis for ranking of the lower and upper quintiles. The undisputable disadvantage of a comparable scheme is that individual results will only be available after the trial has been closed, which could last several months. The system which has been implemented in Germany generates information on the individual performance immediately within 2–3 weeks which is a sufficiently brief period to enable control or revision of methods in an individual laboratory in the case of low performance. In addition, test material can be ordered repeatedly during a 10-month period during which the trial is open for participation so that a short-term repeat in the case of failure and subsequent modification of laboratory methods is feasible.

In conclusion, external proficiency testing as described here fulfils 2 different functions which have to be considered with regard to selection of challenges and composition of test samples in the TMA and also with regard to the terms of evaluation. First, it provides information about the current status of laboratory performance in pathology which is of interest to the collaborating clinician. This information should be based on a representative selection of cases resembling everyday practice. Second, it enables training and improvement of laboratory performance. In order to achieve the latter positive effect, the challenges within the TMA have to be enriched for difficult and borderline cases with low steroid hormone recep-

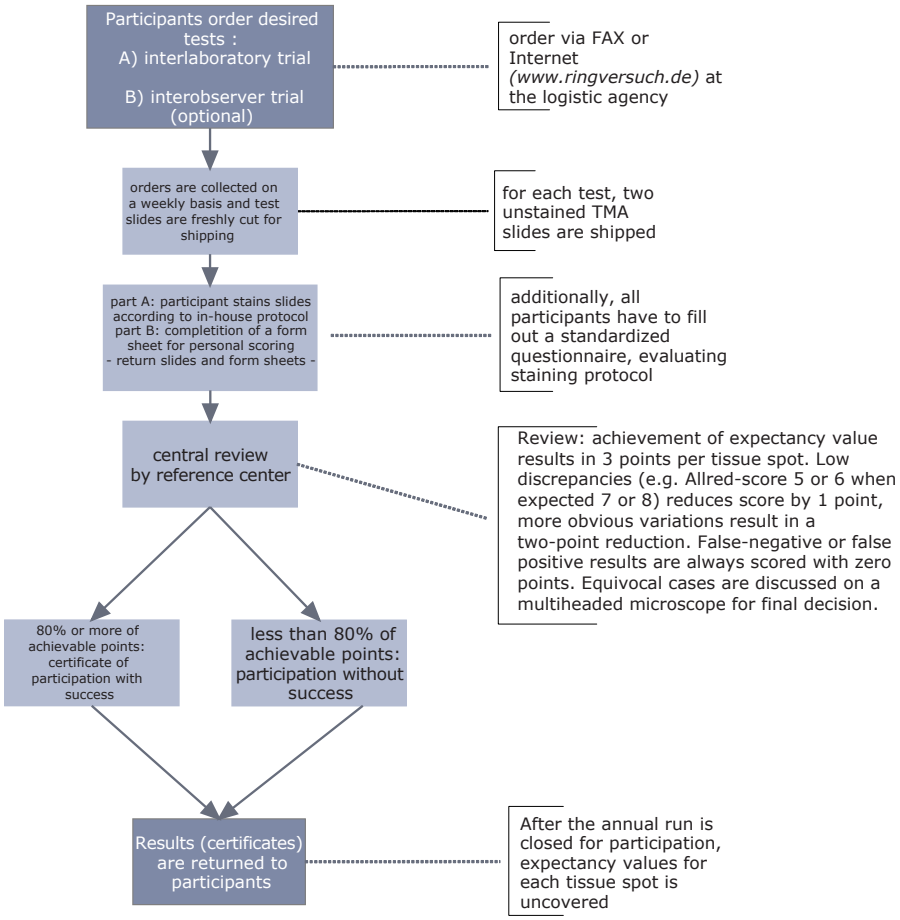

Fig. 3. Organisation and evaluation of the QuIP trial. All immunostains by participants will undergo a central review in order to assess interlaboratory variability. For evaluation a 4-tiered score is applied. Complete agreement with expected result will score 3 points, lesser deviation 2 points, more severe deviation 1 point, and false positivity or negativity 0 points.

tor expression or HER2 2+ status. Because both aims antagonise each other, TMA for interlaboratory trials should be composed of 2 sets of cases which should be evaluated and communicated separately. Accordingly, in future trials there will be a training set and a test set of challenges. Benchmarks to categorise the results on the latter type of challenges will be developed.

Standardisation Requires Further Efforts

Interlaboratory trials may be necessary, but they are not sufficient to assure reproducibility of IHC and FISH. Additional controls have to be included and performed, such as on-slide controls. The latter can be achieved with cell lines embedded in paraffin and sliced like ordinary tissue sections. Cell lines are preferable to tissue samples because a defined content of target can be attributed to individual cell lines. On-slide controls enable correct evaluation of immunostains even when slides are retrieved from the archive. Furthermore, clinicians and pathologists have to collaborate in order to ensure that adequate and rapid fixation of cancer tissue samples used for target analysis will take place according to standardised procedures. Tissue core biopsies from breast cancer are far better suited for a standardised fixation than traditional resection specimens because they are of a uniform size and can immediately be immersed in fixative with rapid and complete diffusion. Adequate fixation requires a minimum period of at least 6 h [3] which should not be shortened. Core biopsies should be processed in a standardised fashion whereby speed and short turn-around time may interfere with quality of immunohistochemical biomarker assays. If adequate fixation period and standardised tissue processing are performed, discrepancies with regard to HER2 scores in tissue core biopsies and resection specimens do generally not occur (>95% identity). Because core biopsies enable standardised fixation, they should be the first source to establish biomarkers in breast cancer, irrespective of potential tumour heterogeneity.

References

1 Cheson BD: Monoclonal antibody therapy for B-cell malignancies. Semin Oncol 2006;33(suppl 5): S2–14.

2 Taylor CR, Levenson RM: Quantification of immunohistochemistry – issues concerning methods, utility and semiquantitative assessment II. Histopathology 2006;49:411–24.

3 Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, Dowsett M, Fitzgibbons PL, Hanna WM, Langer A, McShane LM, Paik S, Pegram MD, Perez EA, Press MF, Rhodes A, Sturgeon C, Taube SE, Tubbs R, Vance GH, van de Vijver M, Wheeler TM, Hayes DF; American Society of Clinical Oncology; College of America Pathologists: American Society of Clinical Oncology/ College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. J Clin Oncol 2007; 25:118–45.

4 Ross JS, Symmans WF, Pusztai L, Hortobagyi GN: Standardizing slide-based assays in breast cancer: hormone receptors, HER2, and sentinel lymph nodes. Clin Cancer Res 2007;13:2831–5.

5 Viale G, Regan MM, Maiorano E, Mastropasqua MG, Dell'Orto P, Rasmussen BB, Raffoul J, Neven P, Orosz Z, Braye S, Ohlschlegel C, Thürlimann B, Gelber RD, Castiglione-Gertsch M, Price KN, Goldhirsch A, Gusterson BA, Coates AS: Prognostic and predictive value of centrally reviewed expression of estrogen and progesterone receptors in a randomized trial comparing letrozole and tamoxifen adjuvant therapy for postmenopausal early breast cancer: BIG 1–98. J Clin Oncol 2007;25: 3846–52.

6 Layfield LJ, Goldstein N, Perkinson KR, Proia AD: Interlaboratory variation in results from immunohistochemical assessment of estrogen receptor status. Breast J 2003;9:257–9.

7 Diaz LK, Sneige N: Estrogen receptor analysis for breast cancer: current issues and keys to increasing testing accuracy. Adv Anat Pathol 2005;12:10–19.

8 Goldhirsch A, Glick JH, Gelber RD, Coates AS, Thurlimann B, Senn HJ: Meeting highlights: international expert consensus on the primary therapy of early breast cancer 2005. Ann Oncol 2005;16: 1569–83.

9 Rudiger T, Hofler H, Kreipe HH, Nizze H, Pfeifer U, Stein H, Dallenbach FE, Fischer HP, Mengel M, von Wasielewski R, Muller-Hermelink HK: Quality assurance in immunohistochemistry: results of an interlaboratory trial involving 172 pathologists. Am J Surg Pathol 2002;26:873–82.

10 Rudiger T, Hofler H, Kreipe HH, Nizze H, Pfeifer U, Stein H, Dallenbach E, Fischer HP, Mengel M, Von Wasielewski R, Muller-Hermelink K; German Society for Pathology; Professional Association of German Pathologists: [Interlaboratory trial 2000 'Immunohistochemistry' of the German Society for Pathology and the Professional Association of German Pathologists]. Pathologe 2003;24:70–8.

11 Rhodes A, Jasani B, Barnes DM, Bobrow LG, Miller KD: Reliability of immunohistochemical demonstration of oestrogen receptors in routine practice: interlaboratory variance in the sensitivity of detection and evaluation of scoring systems. J Clin Pathol 2000;53:125–30.

12 Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, Dowsett M, Fitzgibbons PL, Hanna WM, Langer A, McShane LM, Paik S, Pegram MD, Perez EA, Press MF, Rhodes A, Sturgeon C, Taube SE, Tubbs R, Vance GH, van de Vijver M, Wheeler TM, Hayes DF: American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer. Arch Pathol Lab Med 2007;131:18.

13 Harvey JM, Clark GM, Osborne CK, Allred DC: Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. J Clin Oncol 1999;17:1474–81.

14 Von Wasielewski R, Mengel M, Wiese B, Rüdiger T, Müller-Hermelink HK, Kreipe H: Tissue array technology for testing interlaboratory and interobserver reproducibility of immunohistochemical estrogen receptor analysis in a large multicenter trial. Am J Clin Pathol 2002;118:675–82.

15 Fitzgibbons PL, Murphy DA, Dorfman DM, Roche PC, Tubbs RR: Interlaboratory comparison of immunohistochemical testing for HER2: results of the 2004 and 2005 College of American Pathologists HER2 Immunohistochemistry Tissue Microarray Survey. Arch Pathol Lab Med 2006;130:1440–5.

16 Wells CA, Sloane JP, Coleman D, Munt C, Amendoeira I, Apostolikas N, Bellocq JP, Bianchi S, Boecker W, Bussolati G, Connolly CE, Dervan P, Drijkoningen M, Ellis IO, Elston CW, Eusebi V, Faverly D, Heikkila P, Holland R, Jacquemier J, Lacerda M, Martinez-Penuela J, De Miguel C, Peterse JL, Rank F, Reiner A, Saksela E, Sigal-Zafrani B, Sylvan M, Borisch B, Cserni G, Decker T, Kerner H, Kulka J, Regitnig P, Sapino A, Tanous AM, Thorstenson S, Zozaya E: Consistency of staining and reporting of oestrogen receptor immunocytochemistry within the European Union – an interlaboratory study. Virchows Arch 2004;445: 119–28.