Virchows Arch (2008) 453:537–543
DOI 10.1007/s00428-008-0688-4

ORIGINAL ARTICLE

Proficiency testing of immunohistochemical biomarker assays in breast cancer

Reinhard von Wasielewski · Svenja Hasselmann · Josef Rüschoff · Annette Fisseler-Eckhoff · Hans Kreipe

Received: 30 June 2008 / Revised: 2 September 2008 / Accepted: 7 October 2008 / Published online: 29 October 2008

Abstract Steroid hormone receptor expression and HER2 status have become an integral part of histopathologic characterization of breast cancer and corresponding biomarker assays have gained important prognostic and predictive impact. Because testing inaccuracy could provide a major hazard to modern breast cancer therapy, a laboratory proficiency testing program has been implemented in Germany using tissue microarrays (TMAs). In four consecutive annual trials with 142 laboratories participating on average per trial, estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (Her2) were determined immunohistochemically by participating laboratories followed by central review of all immunostains. Performance strongly depended on the ambiguity of expression of the target molecule in the test samples. In clearly positive (Allred score 7–8; Her2 3+) or negative tissue samples, the majority of participants (86%) achieved concordance rates exceeding 85%. By contrast, low expression of ER or PR (Allred score 3–4) as well as Her2 status 2+ led to considerable lower concordance rates ranging from 41% (Her2 2+) to 75% (PR). Poor reproducibility was predominantly due to inadequate laboratory performance whereas interobserver agreement (weighted kappa statistics) usually was high (>0.81). Laboratories that participated in more than one of the four subsequent trials (n=110) showed a highly significant improvement of performance. In conclusion, a TMA-based proficiency testing of biomarkers in breast cancer has been implemented in Germany over a 5-year period and revealed reliable assessment of unambiguously positive and negative test samples. Low-expressing tumor samples with regard to steroid hormone receptor expression and Her2 status 2+ led to inaccurate evaluations by up to 59% of participants. Regularly participating laboratories showed a significant improvement of performance.

Keywords Steroid receptor · HER2 · Immunohistochemistry · Quality assurance

Reinhard von Wasielewski and Svenja Hasselmann contributed equally to this work.

J. Rüschoff
Institut für Pathologie, Städtisches Klinikum Kassel, Kassel, Germany

A. Fisseler-Eckhoff
Institut für Pathologie, Horst-Schmidt Kliniken, Wiesbaden, Germany

R. v. Wasielewski · S. Hasselmann · H. Kreipe (✉)
Institute of Pathology, Medizinische Hochschule Hannover, Carl Neuberg Strasse 1, 30625 Hannover, Germany
e-mail: Kreipe.Hans@mh-hannover.de

Introduction

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 [1]. In recent years, the human epidermal growth factor receptor 2 (Her2) has gained a 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 [2, 3]. For several reasons both biomarkers usually are determined by pathologists applying tissue sections and immunohistochemistry (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 histopathologic evaluation because they include quantitative assessments whereas the unquestioned strength of histopathology lies in qualitative analysis. In order to cope with the new challenge of target molecule detection in the age of personalized 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 fluorescence in situ hybridization (FISH) and it has been estimated that approximately 20% of current Her2 testing may be false [4]. There is widespread concern that inaccuracy in detection methods and interpretation may lead to an unacceptably high error rate in determining the true hormone receptor status [5]. Comparison of centrally versus locally assessed estrogen receptor (ER) and progesterone receptor (PR) revealed divergent results in a substantial proportion of patients [6]. Obviously, there is a great need to standardize immunohistochemical biomarker assays to further ensure that similar results are obtained by different institutions. External proficiency testing has been proposed as one potential instrument to enable accurate biomarker determination in a noncentralized approach [7, 8]. Yet the most effective setting for external proficiency testing has not been determined. 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.

In this report, the implementation of a nationwide external proficiency testing of ER, PR, and Her2 assessment during five consecutive years from 2002 to 2006 in Germany is described. Unlike previously reported trials [7, 8], tissue microarrays (TMAs) were applied [9, 10] and all immunohistochemical stains done by participants underwent central review in order to enable assessment of interlaboratory and interobserver concordance.

Materials and methods

During the years 2002–2006, four TMAs were generated and distributed to participating laboratories on demand. Tissue cores from routine surgical pathology samples retrieved from the archives of three institutes of pathology in Germany (Hannover, 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) [11] and equivocal positivity for Her2 (2+). The Allred score combines three grades of staining intensity with five percentage categories (proportion of labeled cells). Thus, a sum results with a maximum value of 8 and a threshold for positivity of ≥3 (e.g., intermediate staining intensity of 0–1% cells leads to 2 points (pts) plus 1 pt=3) [11]. Besides equivocal cases, clearly positive or negative samples were included. Only samples that received identical testing in all three laboratories mentioned above entered the final trial. The methods used by the reference laboratories are described in Table 1. From the 2003 to the 2006 run, all test cases with Her2 status of 2+ and 3+ underwent fluorescence in situ hybridization. Among the 3+ tissue samples, 100% revealed amplification of Her2. Among the 2+ tissue samples, 33–56% were polysome and up to 14% (2004, 2006) displayed amplified copy numbers of the Her2 gene. Between 20 and 24 samples were included in the TMAs which were generated exactly as

Table 1 Immunohistochemical methods applied by the reference laboratories

Reference laboratory	Methods	ER	PR	Her2
1	Retrieval method	Pressure cooker at 125°C, 5 min, Citrat buffer pH 6	Pressure cooker at 125°C, 5 min, Citrat buffer pH 6	Citrat buffer pH 6, water bath 95–99°C, 40 min
	Primary antibody	Clone 1D5 und ER-2-123	Clone PGR1294	Rabbit antihuman Her2 protein
	Detection system	PharmDX, DAKO	PharmDX, DAKO	Hercep-test DAKO
	Automat	DAKO Autostainer Plus	DAKO Autostainer Plus	DAKO Autostainer Plus
2	Retrieval method	Specific heat-based antigen retrieval	Specific heat-based antigen retrieval	Specific heat-based antigen retrieval
	Primary antibody	Clone 6F11	Clone PR312	Clone SP3
	Detection system	XT UltraView DAB	XT UltraView DAB	UltraView Universal DAB
	Automat	Ventana Benchmark XT	Ventana Benchmark XT	Ventana Benchmark XT
3	Retrieval method	Pressure cooker at 125°C, 3 min, Citrat buffer pH 6	Pressure cooker at 125°C, 3 min, Citrat buffer pH 6	Pressure cooker at 125°C, 3 min, Citrat buffer pH 6
	Primary antibody	Clone SP1	Clone PR636	Polyclonal rabbit antibody (NCL-cerbB-2p)
	Detection system	ZytoChem Plus HRP Polymer Kit (mouse/rabbit)	ZytoChem Plus HRP Polymer Kit (mouse/rabbit)	ZytoChem Plus HRP Polymer Kit (mouse/rabbit)
	Automat	Manual	Manual	Manual

described [9]. Pathology departments volunteering to participate in external proficiency testing could order up to three slides which were freshly cut and shipped unstained. Within 2 months, immunohistochemical stainings had to be performed and a protocol of the assessment as well as the stained slides had to be returned to the organizers of the trial. Participants were free to perform only one of the three tests or all of them. Unstained slides could be ordered during a 10-month-long period during which the trial was open for participation.

The composition of the TMAs used as test material for all three biomarkers from 2003 to 2006 is depicted in Table 2.

Evaluation

For ER and PR, the Allred score was recorded by the participants and reviewers. For statistical analysis, this score was further simplified likewise with previous studies into four categories: negative cases (Allred 0, 2) or low- (Allred 3, 4), medium- (Allred 5, 6), and high-expressing cases (Allred 7, 8). All scorings and statistics shown are based on this four-tier classification. Each tissue spot was scored. When the expected staining result was achieved or nearly achieved by the participating laboratory, a score of three points was given. In highly steroid hormone receptor positive cases, participants' staining results corresponding to Allred values 7 or 8 were scored 3 pts; Allred values 5 or 6 were scored 2 pts; Allred values 3 or 4 were scored 1 pt; and a negative result was scored 0 points. The scoring system applied for Her2 discriminated between 0 (no staining or membrane staining in less than 10% of invasive tumor cells), 1+ (faint and partial membrane staining in more than 10% of invasive tumor cells), 2+ (weak to moderate complete membrane staining in more than 10% of invasive tumor cells), and 3+ (strong complete membrane staining in more than 10% of invasive tumor cells) [2]. According to steroid hormone receptor evaluation, participants' staining results were reevaluated and accuracy was graded in a 0–3 point system. In case of a Her2 3+ challenge, 3+ was scored with 3 pts, 2+ with 2 pts, 1+ and 0 with 0 point, respectively. With regard to equivocal Her2 cases (2+), 3+ was scored with 1 pt, 2+ with 3 pts, 1+ and 0 with zero points, respectively. False-positive or -negative results were graded as 0 point with regard to steroid hormone receptors as well as Her2. Lost tissue spots (floaters) or tissue fields with obvious technical problems were excluded from evaluation and did not influence the results. According to this procedure, for each participant and marker tested, the maximum sum of achievable points for a specific slide was determined (e.g., 22 tissue spots, one floater, one spot at the edge wiped off: 20 evaluable spots, maximum point score $3\times20=60=100\%$). Based on this information, the achieved sum of points per slide could be transferred into a percentage score. A result of 80% or more was regarded as successful participation (0–59% poor; 60–69% unsatisfactory; 70–79% moderate; 80–89% good; 90–100% excellent). The results are shown as overall performance per marker tested. Moreover, we subdivided the analysis into the different groups of expected results (low-, medium-, and high-expressing cases) to enable a more profound insight into staining capabilities.

Interobserver agreement was assessed using kappa statistic, which takes into account the agreement expected solely on the basis of chance and can be used if more than two categories are classified. Total agreement is indicated by a value of 1.0, but agreement by chance only results in a 0 value.

Although there is no generally accepted value of kappa in the literature that indicates sufficient (i.e., good) agreement, we applied the following guidelines: kappa$<$0.4 represents poor-to-fair agreement; 0.4–0.6 moderate agreement; 0.6–0.8 substantial agreement; and $>$0.8 almost perfect agreement. To measure the grade of agreement, a weighted kappa statistic was performed using the statistic software package SAS, version 8.0.

A standardized questionnaire with routing questions about protocol procedures as well as the participants' personal opinion about the program, problems, and potential improvements was included.

Results

From about 400 laboratories of pathology in Germany, an average of 142 participated in the four proficiency tests

Table 2 Composition of tissue microarrays (percent of tissue spots in TMA, %)

	ER low	ER medium	ER high	PR low	PR medium	PR high	Her2 0/1+	Her2 2+	Her2 3+
2003	26	33	19	31	19	17	52	7	40
2004	26	26	26	11	21	37	36	36	28
2005	14	29	43	19	33	33	33	33	33
2006	27	23	27	27	23	27	36	32	32

ER- and PR-negative samples are not included and will add to 100%

which took place between 2002 and 2006 (Table 1). A total of 12,411 immunohistochemical ER stains underwent central review. In case of PR and Her2, the number was a little bit smaller. A comparison between central review and evaluation by participants revealed an excellent concordance with very low interobserver variability. Interobserver agreement was calculated for 86 participants. Weighted kappa statistics showed a very good overall concordance between participating pathologist and reviewer for all three markers tested: ER 0.84, PR 0.81, and HER2 0.86. These results did not differ from previous trials and did not vary between the different expectancy groups of low-, medium-, and high-expressing cases (detailed data not shown). Therefore, the interlaboratory differences in staining efficiency were the major cause of discordant results. Consequently, the quality of staining and not the reading emerged as the decisive item to be controlled by external proficiency testing. In order to give participants information on the performance of their immunohistochemical detection methods, all analyses were based on the results of the central review of immunostains.

The unequivocally positive samples were detected with high fidelity by participants. Allred scores 7 and 8 were reproduced by 86% of participating laboratories with regard to ER ad PR. More than 92% of participating laboratories achieved correct results with regard to Her2 3+ cases. By contrast, the equivocal cases of Her2 were stained with considerable variation and only 41% of participants scored correctly. Low-expressing ER and PR samples were correctly identified by 61% and 75%, respectively.

The detailed results of all three markers, subdivided into low-, medium-, and high-expressing cases, are depicted in Figs. 1, 2, and 3. Furthermore, the figures display the results for the runs 2002–2003, 2004, 2005, and 2006 sorted by expressing groups. Whereas during the first two runs the rate of poor performers (<60%) was as high as 40% in the group of low-expressing ER cases, this rate has declined in 2005 and 2006. A similar effect could be observed for PR and HER2 2+ cases. With regard to ER and PR, the number of laboratories showing good and excellent results increased from 2002 to 2006. For HER2, false-positivity rates as high as during the first runs (2002, 2004) were reduced in 2005 and disappeared in 2006. The 0/1+ category and the 3+ category in HER2 were correctly diagnosed by the majority of participating laboratories (Figs. 1, 2, and 3.)

Duplicate or triplicate participation in the four subsequent trials was performed by 110 laboratories. A comparison was made based on the percentage score per marker for all cases (overall performance) and for the difficult cases encompassing low ER- and PR-expressing cases and the HER2 2+ category. With regard to ER (all cases), participants obtained significantly better results in the second trial when compared to what they had achieved

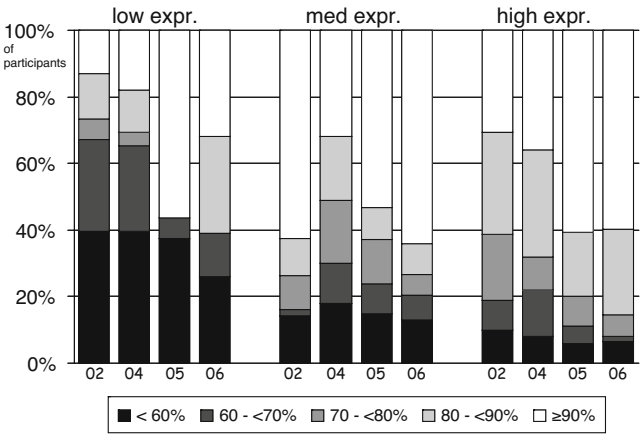

Fig. 1 Results of estrogen receptor staining in four consecutive trials from 2002 to 2006 (*vertical columns*), subdivided according to the level of expression into low, medium, and highly immunopositive cases. The proportion of participants (%) achieving different levels of concordance with expected results is indicated on the *left*. Each *column* encompasses 100% of participants. The *black segments* of the column indicates the proportion of participants with poor concordance (<60%) whereas the *white segment* represents the proportion of participants with good performance (≥90%). Intermediate concordance results (60–<90%) are indicated by different *gray tones* in order to illustrate the difficulty and consequences of setting different thresholds for successful participation. Negative cases had concordance rates of >95% and are not depicted. In all runs from 2002 to 2003, medium- and high-expressing cases had similar results with regards to the good-performance group (>90%, *white segment*) whereas the proportion of participants with poor concordance rates (<60%, *black segments*) were always lowest in the high-expressing group

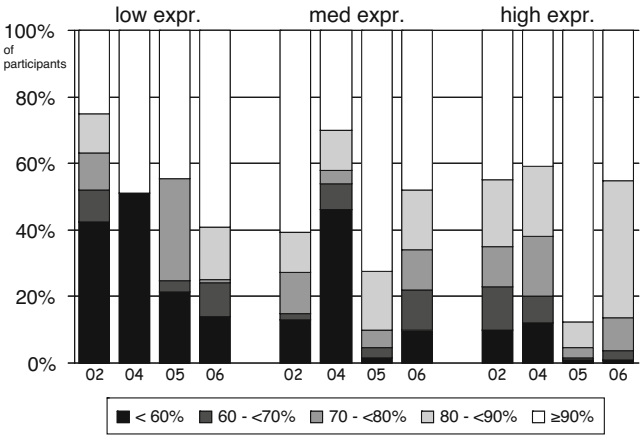

Fig. 2 Results of progesterone receptor staining in four consecutive trials from 2003 to 2006 (*vertical columns*), subdivided according to the level of expression into low, medium, and highly immunopositive cases. The proportion of participants (%) achieving different levels of concordance with expected results is indicated on the *left*. Each *column* encompasses 100% of participants. The *black segments* of the column indicates the proportion of participants with poor concordance (<60%) whereas the *white segment* represents the proportion of participants with good performance (≥90%). Intermediate concordance results (60–<90%) are indicated by different *gray tones* in order to illustrate the difficulty and consequences of setting different thresholds for successful participation. Negative cases had concordance rates of >95% and are not depicted

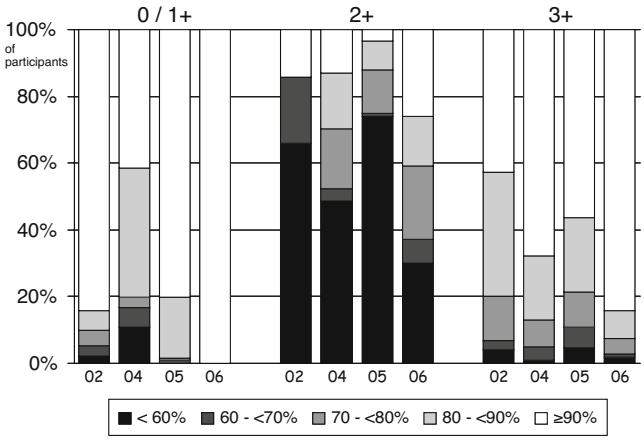

Fig. 3 Results of Her2 staining in four consecutive trials from 2003 to 2006 (*vertical columns*), subdivided according to the level of expression into negative (0, 1+), equivocal (2+), and positive (3+) cases. The proportion of participants (%) achieving different levels of concordance with expected results is indicated on the *left*. Each *column* encompasses 100% of participants. The *black segments* of the column indicates the proportion of participants with poor concordance (<60%) whereas the *white segment* represents the proportion of participants with good performance (≥90%). Intermediate concordance results (60–<90%) are indicated by different *gray tones* in order to illustrate the difficulty and consequences of setting different thresholds for successful participation

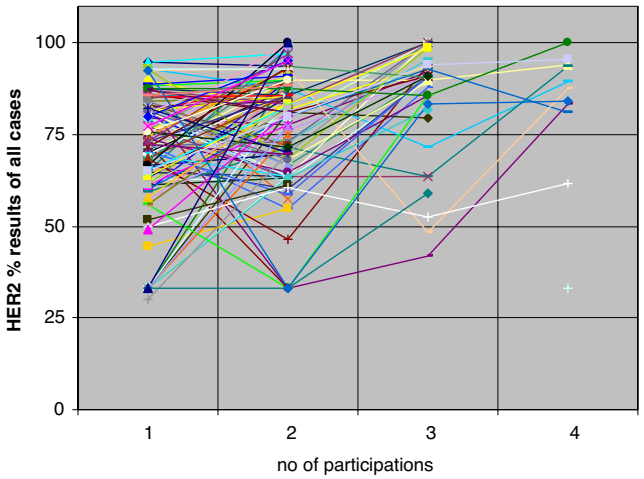

Fig. 4 Percentages of maximum concordance with expected results achieved by individual laboratories in up to four consecutive participations over time. Regardless of the year in which the first participation took place, a duplicate or triplicate participation led to higher concordance scores over time as can be seen from most of the *ascending lines*, each of which represents one single laboratory. Results are shown for HER2 but were similar for ER and PR

before in their first trial (p=0.008; Wilcoxon test). An improvement of performance could also be seen in case of triplicate participation (p<0.001; Friedman test). This improvement could be observed irrespective of which of the four subsequent trials were passed by the laboratories. Hence, the improvement appeared not to be due to different levels of difficulty between the subsequent trials. If low-expressing cases only were included in the analysis, the improvement of the percentage score was 10.8% on average. There was a significant improvement again with regard to duplicate (p=0.008) as well as triplicate participation (p< 0.001). When PR is considered, there was a significant improvement of 7.2% from the first to the second run and to the third run (both p<0.001). With regard to the low-expressing PR cases, the effect was even more pronounced with an improvement of 33.1% (p=0.001 and p=0.003, respectively).

Similarly, the results for HER2 improved when laboratories participated in duplicate or triplicate. When all challenges are considered, a mean improvement of 10.2% was seen in case of duplicate and triplicate participation (p<0.001 and p< 0.001, respectively; Fig. 4). As has been seen with steroid hormone receptors, the improvement of performance with duplicate or triplicate participation became even more obvious when equivocal challenges were taken into consideration. In the 2+ subgroup, there was an increase of 24.9% (p=0.004; one versus two runs) and also when three runs were performed (p<0.001).

Discussion

Breast cancer in recent years has functioned as a pioneer tumor, setting the stage for a new era of diagnostic and therapy in oncology. Steroid hormone receptors and Her2 provided the first examples for targeted therapy and marked the beginning of the age of personalized 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 IHC or FISH. In most countries, the latter methods are predominantly used to assess target molecules in breast cancer [12]. 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 categorization have to be defined [13, 14]. The biological significance and justification of these thresholds is particularly unclear in a gray zone between unequivocal positive and negative cases [1, 4, 15]. Whether this gray zone could be diminished if the corresponding assays were more reliable is a matter of debate [15]. 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 [6, 15, 16]. Despite these potential hazards, IHC offers a number of decisive advantages like correlation to number of tumor cells and their viability as well as to admixture of normal, noninvasive, 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. External proficiency testing is a useful tool to provide this information [4] and national external quality assessment schemes for immunohistochemistry like NEQUAS-ICC have been founded [7, 8]. Whereas the potential benefit of quality assurance trials is unquestioned, there are many open questions with regard to composition of test samples and evaluation. In this report, we have described the German approach and the experiences derived thereof. In order to increase the number of challenges and to diminish the variability of individual tumors or modes of fixation, we have applied TMA with tissues from three different institutions. The usefulness of TMA in interlaboratory trials has been demonstrated before [9, 10, 17]. The compliance of participants with the TMA was very good and reading of the immunostained spots was not a cause of trouble. As has been described previously, the overall concordance was high in unequivocally positive or negative cases [8]. Discordant results with a high percentage of false-negative scorings were encountered in the low steroid hormone receptor positive group and Her2 2+ cases. In each of the four trials, these kinds of borderline cases were enriched in the TMA in order to provide effectively discriminatory challenges [17]. This overrepresentation of difficult cases led to a higher proportion of underperforming laboratories than would have been expected with a more representative composition of challenges in the TMA. Because no generally accepted benchmark criteria are available which could be adapted to the level of difficulty and arbitrary bench marks have to be set, the results are prone to misunderstanding when communicated to nonpathologists, e.g., clinicians. On the other hand, the TMAs enriched for difficult borderline cases will more effectively alert pathologists to potential shortcomings in their immunohistochemistry laboratories. Indeed, it became evident from the analysis of laboratories participating in duplicate or triplicate that the performance significantly improved. These findings demonstrate that, besides testing, external proficiency trials exert a training function and improve overall performance [17]. Furthermore, participants benefited from the information derived from the questionnaires as it became evident which steps of the immunohistochemical staining procedure are particularly crucial. For example, heat pretreatment for antigen retrieval was found to be very heterogeneous with microwaving being overrepresented in the low-performance group. Again, these differences became effectively visible when the difficult challenges were considered. The obvious relevance of antigen retrieval is particularly suited to demonstrate a dilemma which proficiency testing programs in pathology are facing. Whereas standardized material, e.g., cell lines, with known content of the target protein in question would be ideal to evaluate and compare sensitivity of detection methods, diagnostic routine pathology is usually challenged with heterogeneous tissues modified by fixation and embedding. Thus, fixed tissue from resection specimens suffers from the drawback of not being standardized but as test material it covers more and relevant steps of the diagnostic process than would be possible with cell lines.

In conclusion, external proficiency testing as described here fulfills two different functions which have to be considered with regard to selection of challenges and composition of test samples in the TMA as well as with regard to the terms of evaluation. First, it provides information about the current status of laboratory performance. 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 receptor expression or Her2 2+ status. Because both aims antagonize each other, TMA for interlaboratory trials should be composed of two sets of cases which should be evaluated and communicated separately. Accordingly, in future trials, there should be a training set and a test set of challenges. Benchmarks to categorize the results on the latter type of challenges need to be developed.

Acknowledgements The authors thank the numerous participants of the trials 2002–2006 for helpful and critical comments given in order to improve quality assurance trials in Germany.

Conflict of interest statement R. von Wasielweski is founder and partial owner of the "Multiblock GmbH" which served as a logistics center in the 2006 trial.

References

1. Goldhirsch A, Glick JH, Gelber RD et al (2005) Meeting highlights: international expert consensus on the primary therapy of early breast cancer 2005. Ann Oncol 16:1569–1583
2. Slamon DJ, Leyland-Jones B, Shak S et al (2001) Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med 344:783–792
3. Carlson RW, Moench SJ, Hammond ME et al (2006) NCCN HER2 testing in Breast Cancer Task Force. HER2 testing in breast cancer: NCCN Task Force report and recommendations. J Natl Compr Canc Netw Suppl 3:S1–S22
4. Wolff AC, Hammond ME, Schwartz JN et al (2007) 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 25:118–145

5. Ross JS, Symmans WF, Pusztai L et al (2007) Standardizing slide-based assays in breast cancer: hormone receptors, HER2, and sentinel lymph nodes. Clin Cancer Res 2007, 13:2831–2835
6. Viale G, Regan MM, Maiorano E et al (2007) 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 25:3846–3852
7. Rhodes A, Jasani B, Barnes DM et al (2000) 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 53:125–130
8. Wells CA, Sloane JP, Coleman D et al (2004) Consistency of staining and reporting of oestrogen receptor immunocytochemistry within the European Union—an interlaboratory study. Virchows Arch 445:119–128
9. von Wasielewski R, Mengel M, Wiese B et al (2002) Tissue array technology for testing interlaboratory and interobserver reproducibility of immunohistochemical estrogen receptor analysis in a large multicenter trial. Am J Clin Pathol 118:675–682
10. Hsu FD, Nielsen TO, Alkushi A et al (2002) Tissue microarrays are an effective quality assurance tool for diagnostic immunohistochemistry. Mod Pathol 15:1374–1380
11. Allred DC, Harvey JM, Berardo M et al (1998) Prognostic and predictive factors in breast cancer by immunohistochemical analysis. Mod Pathol 11:155–168
12. Harvey JM, Clark GM, Osborne CK et al (1999) 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 17:1474–1481
13. Taylor CR, Levenson RM (2006) Quantification of immunohistochemistry—issues concerning methods, utility and semiquantitative assessment II. Histopathology 49:411–424
14. Swanson PE, Schmidt RA (2005) Beneath the surface of the mud, part II: the dichotomization of continuous biologic variables by maximizing immunohistochemical method sensitivity. Am J Clin Pathol 123:9–12
15. Diaz LK, Sneige N (2005) Estrogen receptor analysis for breast cancer: current issues and keys to increasing testing accuracy. Adv Anat Pathol 12:10–19
16. Layfield LJ, Goldstein N, Perkinson KR et al (2003) Interlaboratory variation in results from immunohistochemical assessment of estrogen receptor status. Breast J 9:257–259
17. Fitzgibbons PL, Murphy DA, Dorfman DM et al (2006) 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 130:1440–1445