External Quality Assurance of HER2 FISH and ISH Testing

Three Years of the UK National External Quality Assurance Scheme

John M.S. Bartlett, PhD, FRCPath,1,2 Merdol Ibrahim, PhD,2 Bharat Jasani, PhD, FRCPath,3 John M. Morgan,3 Ian Ellis, BM, FRCPath,4 Elaine Kay, FRCSI, FRCPath,5 Yvonne Connolly, MSc,6 Fiona Campbell, MSc, FIBMS,1 Anthony O’Grady, MSc, PhD,5 Sarah Barnett, MSc,7 and Keith Miller, FIBMS2

Key Words: HER2; Fluorescence in situ hybridization; In situ hybridization; Quality assurance; Practice guidelines; Gene/chromosome ratio; Gene copy number

DOI: 10.1309/AJCPLN78ZQXEMNMA

Abstract

The American Society of Clinical Oncology/College of American Pathologists guidelines highlighted the critical importance of quality assurance in diagnostic testing for HER2.

Unstained formalin-fixed, paraffin-embedded human breast carcinoma cell line sections were circulated to scheme participants on 9 occasions. “Reference laboratories” reported results for the HER2/chromosome 17 ratio and HER2 copy number for 3 years for each cell line, including 418 sets of results (1,671 results total). The number of participants was 62 laboratories in the final analysis.

The mean and SD of results from reference laboratories demonstrated consistency during the 3-year period. The percentage of laboratories achieving “appropriate” results ranged from 45% to 88%, and the percentage achieving “inappropriate” results ranged from 5% to 29%. No consistent effect of the HER2 in situ hybridization testing method was demonstrated.

Participation in external quality assurance schemes is a valuable mechanism for demonstrating and acquiring consistency for HER2 testing by in situ hybridization. Poor performance can be corrected via assistance and advice.

The recent publication of the American Society of Clinical Oncology/College of American Pathologists (ASCO-CAP) guidelines for HER2 testing1 reignited the debate about quality assurance in diagnostic testing for the HER2 gene and its protein product. Clearly, if benefits reported for patients receiving trastuzumab (Herceptin) in trials2,3 are to be translated into daily management of breast cancer, accurate diagnostic approaches to assess the HER2 status of breast cancers are critical.

In situ hybridization techniques, colorimetric (CISH) and fluorescence (FISH), have been widely applied in breast cancer and have confirmed that testing for HER2 gene amplification is a valid predictor of HER2 expression and response to trastuzumab.4-10 We previously reported on the pilot phase of the United Kingdom National External Quality Assessment Service (NEQAS),11,12 and others have reported the consistency of FISH diagnostics in large reference laboratory studies13 and in “ring studies.”14 However, once again, the quality of HER2 FISH testing done outside so-called reference centers has been questioned.

The UK NEQAS has maintained its multinational scheme to assess the technical quality of laboratories using HER2 FISH testing for 3 years (details can be found at http://www.UKNEQASicc.ucl.ac.uk). Data from 6 reference laboratories were used to define acceptable criteria for HER2 FISH testing based on a panel of 4 cell lines tested in duplicate. Results from 62 active participants from 11 countries (United Kingdom, Austria, Denmark, Republic of Ireland, France, Norway, Portugal, Slovenia, Switzerland, the Netherlands, and Hong Kong) are now available covering a period of 36 months. In addition, we have now collated data reflecting the relative performance of the major diagnostic assay systems.
Materials and Methods

Unstained sections from formalin-fixed, paraffin-processed blocks of human breast carcinoma cell lines—MDA-MD-231, MDA-MB-453, BT-20/MDA-MB-175 (the BT-20 cell line was replaced in year 1 and only the MDA-MB-175 cell line was used thereafter), and SKBR3—were circulated to participants in the scheme on 9 separate occasions during a 36-month period. We have previously reported on data from the first 12-month period of the scheme showing the SKBR3 and MDA-MB-453 (this cell line can be amplified or borderline) cell lines to have HER-2/neu gene amplification, whereas the cell lines BT-20, MDA-MB-175, and MDA-MB-231 lines were negative. For the study, 4-µm sections, cut from paraffin-embedded cell pellets provided by Novocastra, Newcastle upon Tyne, England, were assayed by different FISH methods in the reference laboratories (PathVysion [Vysis, Downers Grove, IL], PharmDx [Dako, Cambridge, England], or Ventana Inform [Ventana, Tucson, AZ]). Cell lines were selected to span the critical diagnostic region of low-level amplification (2-3 copies per chromosome). The reference laboratories returned results for each cell line relating to the HER2/chromosome 17 ratio and HER2 copy number (to assess laboratories using HER2 copy number only). Results from reference laboratories and participants were then assessed for each cell line and over time.

Results

Figure 1 summarizes the results from reference laboratories (5 using the HER2/chromosome 17 ratio and 1 using HER2 copy number) for each cell line in 9 separate runs. The mean and SD of results obtained from reference laboratories for each cell line demonstrated consistency of the reference laboratories in determining the HER2 status of these cell lines during a 3-year period (Figure 1). The mean (± SD) HER2 ratios for all 9 runs reported by reference laboratories for the individual cell lines were as follows: SKBR3, 3.67 ± 0.22; MDA-MB-453, 2.22 ± 0.14; MDA-MB-175, 1.20 ± 0.12; and MDA-MB-231, 1.10 ± 0.4. The mean variation in results from reference laboratories and participants increased from 33 and 30, respectively, in run 1 to 67 and 62, respectively, in run 9 (each participant received 2 test slides; some returned both sets, leading to multiple submissions). Data presented herein are, therefore, from a total of 1,571 analyses of these 5 cell lines performed by 62 separate laboratories during a 3-year period.

In total, 88 sets of results were returned from the reference laboratories for the 9 runs (an average of 1.6 results per laboratory per run), representing 351 FISH results. For the study, 395 slide sets were distributed to participating laboratories during the 9 runs covered by this analysis. Of the 395 sets, 330 sets of data (83.5%) were returned from participants, representing 1,320 HER2 FISH results submitted from participating laboratories (excluding the 351 results submitted by the reference laboratories). The number of submissions and participants increased from 33 and 30, respectively, in run 1 to 67 and 62, respectively, in run 9 (each participant received 2 test slides; some returned both sets, leading to multiple submissions). Data presented herein are, therefore, from a total of 1,571 analyses of these 5 cell lines performed by 62 separate laboratories during a 3-year period.

Figure 1 Mean ± SD (error bars) for the HER2/chromosome 17 ratio (A) and HER2 copy number (B) results for 9 external quality assurance runs from reference laboratories. 0 represents average results from all 9 runs. Individual run results are then plotted for the 3+ (circle), 2+ (square), 1+ (diamond), and 0+ (triangle) National External Quality Assessment Scheme cell lines. The mean (± SD) HER2 ratios for all 9 runs reported by reference laboratories for the individual cell lines were as follows: SKBR3, 3.67 ± 0.22; MDA-MB-453, 2.22 ± 0.14; MDA-MB-175, 1.20 ± 0.12; and MDA-MB-231, 1.10 ± 0.4. The mean variation in results for HER2 ratios for the 9 runs reported by reference laboratories was 9.9%.
results for HER2 ratios for the 9 runs reported by reference laboratories was 9.9%. These data demonstrate that results were maintained within a narrow band for each cell line for the 9 assessments for determination of the HER2 copy and the HER2/chromosome 17 ratio.

The reference laboratory results were used to calculate an acceptable range (based on copy number and gene/chromosome ratio) as previously described. Briefly, participant results within the range of reference laboratories were regarded as “appropriate.” Data from the reference laboratories suggest that the “appropriate” range represents a range of 10% around the estimated mean value for each cell line in each run. Results within 10% of the lower or upper limits of reference results were defined as “acceptable” based on previous data from our laboratories on interobserver variation in FISH, which is also reflected in the recent ASCO-CAP guidelines.

Data from other laboratories on observer variation for the estimation of quantitative HER2 ratios are lacking, although concordance is consistently reported as high (eg, Press et al). Results outside these limits were regarded as “inappropriate.” Notwithstanding these definitions, any sample that resulted in a misdiagnosis (eg, a nonamplified sample scored as amplified) was regarded as inappropriate.

By using these criteria, participants’ results were scored, and overall performance for the 4 cell lines was categorized as appropriate, acceptable, or inappropriate for each run. If no submission was received, these results were excluded from the current analysis. During the 3-year period of the scheme between 30 (run 1) and 62 (run 9) laboratories subscribed to the scheme. The percentage of laboratories achieving appropriate results, based on those returning results, ranged from 45% (run 4) to 88% (run 6), and the percentage of laboratories achieving inappropriate results ranged from 5% to 29%.

For runs 4 through 9 of the scheme, the impact of the assay method used to determine HER2 status on laboratory performance was assessed. It should be noted that this

**Figure 2.** The performance of 3 individual laboratories is provided for reference showing performance over time; we selected laboratories to show the fluctuations in performance over time, not to show the best cases.

**Figure 3.** Exemplar results from 3 separate laboratories plotted against reference laboratory means (± SD) over time. Individual run results are plotted for the 3+ (circle), 2+ (triangle), 1+ (square), and 0+ (diamond) National External Quality Assessment Service cell lines. A, Laboratory A, HER2/chromosome 17 ratio. B, Laboratory B, HER2 copy number. C, Laboratory C, HER2/chromosome 17 ratio.
analysis is weighted by the majority of participants using the PathVysion system for HER2 FISH testing (27-39 laboratories per run), with fewer using the PharmDx (6-16 laboratories per run) and Inform (6-9 laboratories per run) \( \text{IFigure 4A}\); some laboratories were using CISH tests (chromogenic or silver-enhanced detection), but there were too few results at this stage for meaningful analysis.

All results for runs 4 through 9 were plotted by method used and result obtained \( \text{IFigure 4B}\). The small number of participants using the Inform and PharmDx assays required combination of results over time to assess performance of laboratories using these tests. There were 36 submissions during 6 runs from laboratories using the Inform test, 58 from laboratories using the PharmDx kit, and 176 from laboratories using the PathVysion kit. The number of laboratories using the Inform kit increased only slightly during this period, whereas there was a 45% increase in laboratories using the PathVysion kit and almost a tripling in the number of laboratories using the PharmDx kit. \( \text{IFigure 5}\) reflects the performance of the laboratories using these different methods, relative to results from reference laboratories (using a mix of all 3 methods) for the 6 runs for which data are available. Owing to the bias toward laboratories using the PathVysion kit, numbers are insufficient at this stage for formal statistical analysis.

**Discussion**

HER2 protein expression or gene amplification in breast cancer is a recognized predictive biomarker for response to trastuzumab in early and advanced disease.\(^\text{18}\) Testing procedures follow national guidelines,\(^\text{4,19}\) and, increasingly, the importance of external quality assurance is being recognized. We have previously reported on the UK NEQAS FISH module\(^\text{12}\) during the first 12 months of data collection...
from participants. This scheme is now mandatory for UK laboratories performing FISH and available as a pilot scheme for laboratories using other in situ methods (eg, silver in situ hybridization [SISH] or CISH) for the HER2 oncogene in clinical breast cancers under Clinical Pathology Accreditation (UK) and National Quality assurance assessment program regulations. The scheme is also open to non-UK laboratories as an aid to monitoring and improving performance. The scheme is focused on testing the accuracy of the methodological aspects of ISH testing for HER2 gene amplification. Other schemes, including those run by ASCO-CAP, the UK Royal College of Pathologists, and other national bodies are essential to ensure that pathologic diagnoses (eg, discrimination of invasive from noninvasive tissues and correct identification of pathologic type) are of high quality. Quality assurance is essential for every facet of the diagnostic process, and the UK NEQAS ISH scheme forms part of a spectrum of quality monitoring that is essential for every diagnostic laboratory.

Published reports on the performance of FISH, CISH, and SISH assays for the determination of HER2 status focus predominantly on the relationship between central and local laboratories and performance differences between these different types of facilities. Such reports have been highlighted as showing an alarming degree of inconsistency between test results for immunohistochemical and molecular analyses of HER2 in different settings.1 However, relatively little information is available on the performance of individual diagnostic laboratories relative to known standards and for a prolonged period. The UK NEQAS scheme has now collected data from more than 60 laboratories for a 3-year period to assess the quality of HER2 FISH testing across diagnostic laboratories.

The data we present herein confirm and extend earlier findings2 that reference laboratories, such as those used in the UK NEQAS scheme, can provide consistent, high-quality, and reproducible HER2 analyses using currently available technologies, including the Inform, PharmDx, and PathVysion methods (Figure 1). These data are in line with other literature on “ring” and concordance studies involving such reference laboratories,1,14,20 and support the conclusion that these methods can be adequately “quality assured” to provide consistent and reproducible diagnostic results. However, it is laboratories outside the often self-selecting circle of reference laboratories whose performance is most frequently questioned on the basis of comparisons between results from multiple local laboratories, which are combined for comparison with single central laboratories.13,21 Such analyses may condemn the many for the failures of the few. The UK NEQAS scheme, however, treated each participating laboratory’s results separately and also included an analysis of performance over time. Data from the scheme suggest that the majority of diagnostic laboratories are performing ISH-based diagnostics for HER2 of high quality (Figure 2). Furthermore, there is benefit to be derived from participation in external quality assurance schemes because in the latter phase of the scheme (Figure 2) and among “experienced” participants (defined as those with >6 submissions to the scheme), the quality of testing is demonstrably higher. However, as the monitoring of individual laboratories shows (Figure 3), performance, even in experienced centers, may dip and should be regularly monitored to avoid errors.

There are now multiple diagnostic tests approved for the molecular assessment of HER2 gene amplification, including single- and dual-copy FISH and CISH (including SISH).1,13,22 Collection of information on the methods used by participating laboratories suggests that the vast majority continue to use FISH-based methods, although evidence is emerging that as CISH- and SISH-based methods rapidly improve in ease of use and interpretation, this is likely to change. Even among the FISH-based methods, there is a strong preference for the longest established method, PathVysion. However, significant numbers of participants are now using the PharmDx and Inform tests (Figure 4A).

It is now possible for a preliminary assessment of the impact of these methods on performance. Extreme caution must be exercised in this early analysis, however, since the data are biased by a number of factors. First, there are markedly more laboratories using the PathVysion test than any of the competing assays. This clearly weights any analysis toward this test. Second, because the Inform and PharmDx tests are newer on the market, there is a natural bias toward less experience with these reagents, which may affect the results obtained without providing evidence that these are de facto less accurate methods. Indeed, as mentioned, the number of UK NEQAS participants for the PharmDx kit has nearly tripled during the life of the scheme, with many participants joining in the last year. Conversely, the number of participants using the other methods has risen more steadily. Finally, the data provided for the UK NEQAS scheme are, by necessity, collated from different laboratories; thus, it is dangerous to draw overly strong conclusions from the performance of different methods in different laboratories, which might lead to one test being recommended over another.

Given these caveats, the data presented herein (Figures 4B and 5) are highly encouraging. There is a suggestion that users of the PathVysion system score more highly more frequently than the users of competing assays, but this is not a marked difference (Figure 4B). Indeed, proportionately fewer results from the Inform test were “inappropriate” than with the other 2 tests. Similarly, results obtained for each cell line, for multiple runs in a 2-year period, would seem to suggest that the PathVysion kit provides more stable and reproducible results (Figure 5). However, the significantly larger proportion of results for this test weighs against small numbers of “inappropriate” performers, biasing results, while for the Inform and PharmDx kits, a small proportion of outliers can...
adversely affect these results. For this reason, we have not performed a formal statistical analysis of differences between methods at this stage. Our conclusion at present is that, unlike immunohistochemical tests, there is no strong evidence to suggest that one method performs more robustly than the others. It may be that future evidence of such a difference emerges; however, at this time, we see no requirement to recommend one method over another. CISH and SISH technologies have not yet been evaluated, yet it is likely that these methods will also prove equivalent in terms of accuracy.

The conclusions to be drawn from this analysis of diagnostic laboratory performance are broadly encouraging. Many more laboratories perform well than perform poorly. High-quality HER2 testing is not solely confined to reference laboratories. Participation in the UK NEQAS scheme seems to provide a mechanism for improving performance23,24; however, for a small proportion of centers, further improvements in quality are required. The adoption of the UK NEQAS FISH scheme by the National Quality assurance assessment program will ensure that further support is offered to such centers in the future.

From the 1Endocrine Cancer Group, Edinburgh Cancer Research Centre, Western General Hospital, Edinburgh, Scotland; 2UK National External Quality Assessment Service and 3Advanced Diagnostics, University College London, London, England; 4Department of Pathology, School of Medicine, Cardiff University, Cardiff, Wales; 4University of Nottingham, Molecular Medical Sciences, Department of Histopathology, Nottingham City Hospital, Nottingham, England; 3Department of Histopathology, Beaumont Hospital and the Royal College of Surgeons in Ireland, Beaumont Hospital, Dublin; and 4Adelaide and Meath Hospital, incorporating the National Children’s Hospital, Dublin.

Address reprint requests to Dr Bartlett: Endocrine Cancer Group, Edinburgh Cancer Research Centre, Western General Hospital, Crewe Rd S, Edinburgh EH14 2XR, Scotland.

References


