Author's response to reviews

Title: Inter-observer reproducibility of HER2 immunohistochemical assessment and concordance with fluorescent in situ hybridization (FISH): pathologist assessment compared to quantitative image analysis

Authors:

Gulisa Turashvili (gturashv@bccrc.ca)
Samuel Leung (Sam.Leung@vch.ca)
Dmitry Turbin (dmitry.turbin@vch.ca)
Kelli Montgomery (kelli.montgomery@stanford.edu)
Blake Gilks (Blake.Gilks@vch.ca)
Rob West (rbwest@stanford.edu)
Melinda Miller (memiller@bccancer.bc.ca)
David Huntsman (dhuntsman@bccancer.bc.ca)
Samuel Aparicio (saparicio@bccrc.ca)

Version: 2 Date: 17 February 2009

Author's response to reviews: see over
Cover Letter

February 17th, 2009

RE: “Interobserver reproducibility of HER2 immunohistochemical assessment and concordance with fluorescent in situ hybridization (FISH): pathologist assessment compared to quantitative image analysis” by Gulisa Turashvili, Samuel Leung, Dmitry Turbin, Kelli Montgomery, Blake Gilks, Rob West, Melinda Carrier, David Huntsman, Samuel Aparicio

Dear Editor-in-Chief and Reviewers,

Thank you very much for considering our manuscript for publication in BMC Cancer. We have revised the manuscript according to the reviewers' suggestions. We believe that all comments have been addressed in a satisfactory manner and this revised manuscript now meets the standards for publishing in BMC Cancer. Once more, thank you very much for your interest in our work, and we look forward to hearing from you.

Reviewer: 1

Major compulsory revisions:

1. “First of all the authors should explain more in detail why they only have complete data from 616 patients.”

Of 4,046 cases, the number of cores scorable by all four observers (visual or machine) on IHC slides, regardless of FISH status, was 1212. The number of cases scorable by both IHC and FISH was 616. When scoring IHC slides, exclusion criteria included a variety of reasons such as core drop-off during the processing, insufficient or absent tumor tissue within the cores, or artifactual distortion of the tissue making interpretation impossible. For FISH analysis, in addition to these exclusion criteria, tumors that failed to hybridize or did not allow to count >40 tiles were not included in the analysis. This led to the decreased number of cases with available IHC and FISH data. The revised manuscript now includes data from only 1212 patients as suggested.

2. “Then they should explain how many patients they have for each calculations. When you write ‘whole group’ how many patients are included. My advice would be to use only those patients of whom all data is available so 1212 for the IHC and 616 for the IHC/FISH comparisons.”

The kappa statistics, log-rank tests and their corresponding permutation tests were all based on the maximum number of cases available for each comparison. We re-analyzed
the results using the data from 1212 patients when comparing IHC scores, and the data from 616 patients when comparing FISH and IHC scores. The manuscript has been updated accordingly.

3. “Authors should give an overview of the therapeutical consequences of using a machine to do the scoring does this result in more FISH (2+ cases) assessments or more 1+ false negative cases? How many 2+ cases would have been scored as 1+ or 3+ cases by machine 1 in comparison to an observer. Show results only for the 616 cases with all data.”

To assess the therapeutical consequences of using a machine for HER2 scoring, we summarized visual scores and FISH results in Table 6. It seems that automated scoring on the Ariol machine leads to more 2+ scores (2-3 times as many as visual scoring) and thus it would result in more FISH assessments in clinical practice.

4. “Table 1: FISH non-amplified. Numbers do not add up some have 1074 cases other 1084 and machine 2 only 20 cases. Explain the differences, comparison should only be shown for those with all data available. Why does machine two only have 20 cases?”

Table 1 has been updated and comparisons are only shown for 616 cases, i.e. patients with all data available.

5. “Why does machine 1 result in 506 unscorable cases is this due to the 50 cell limit? If so, this means that both FISH and observers scored HER2 in less than 50 tumor cells? How many cases were included in this study with less than 50 tumor cells? What was the lower limit of number of tumor cells (1, 10, 25) that was accepted for analysis by the observer or for FISH?”

For Ariol scores, all cores with less than 50 cells were considered unscorable. For visual scores, pathologists used clinical approach – cores with obvious negative or 3+ staining were also scored even if the number of tumor cells was between 10-50, based on semiquantitative visual assessment. For FISH, we only accepted scores if >40 tiles were counted. With Metafer system, one tile is considered 1 cell as the size of a tile is approximately the average size of a nucleus.

6. “Measuring FISH by Metasystems in such a tissue will include many normal cells as well, were these tissues scored manually as well?”

Normal cells were excluded wherever possible. We reviewed corresponding H&E slides when needed.

7. “If agreement between system 1 and 2 is 1.00 (95% CI 1-1) for binarized HER2 (Table 2), why is the agreement between observer 1 vs system 1 0.914 and vs system 2 0.667? One should expect these agreements to be similar as well.”
This was caused by different number of cases for each comparison as we tried to use the maximum available cases for each comparison. This may give machine scoring some advantage as the machine was trained to be more specific than sensitive in order to avoid false positives. This may explain the high number of 2+ scores for automated scores. The machine excluded cores with <50 cells from analysis which may be a reason for having so many uninterpretable cases. Using only 1,212 cases with all IHC data available, this discrepancy has been eliminated.

8. “P-values for the Kaplan-Maier curves are amazingly low. Is this the p-value from the log rank? How do these get so low? Her2 is prognostic but not that prognostic. What is the chosen end-point overall survival or breast cancer related death/recurrences?”

The p-values are from log-rank tests. The reason for such low p-values was the large number of cases on the series. The end-point is breast cancer specific death.

9. “Please note the publications by Skaland et al. on Her2 digital assessment.”

This paper has been included in references.

Reviewer: 2

Major compulsory revisions:

1. “(1) Table 1: Results for FISH non-amplified cases (n=1084) contains errors: (a) Visual 1, row (0, 1+, 2+, 3+ and unscorable) adds up to only 1074, not 1084. Should “unscorable” have been 110? (b) Machine 1, row (0, 1+, 2+, 3+ and unscorable) also adds up to only 1074, not 1084. (c) Machine 2, row (0, 1+, 2+, 3+ and unscorable) has only 4 entered for every value, total 20. This calls into question the integrity of the other statistics reported, and the calculations should be included with any revision.”

Table 1 has been updated as we only used data from 1212 patients. All other calculations have also been checked for accuracy.

2. “The level of agreement between the two instruments and between the instruments and the pathologists requires explanation in terms of the observed tissue morphological and staining features, in addition to the statistical discussion. In particular, there is a wide variation in the levels of agreement between the scores from the four different sets of IHC scores and the FISH results. What features were responsible for this? For example, how many specimens exhibited chromosome 17 polysomy? How many showed heterogeneous gene amplification or protein expression in which evaluation limited regions of the core could have played a role? In their discussion the authors need to separate differences in the actual staining and interpretation (level of staining, numbers of spots, etc.) and underlying mechanisms.”
When we scored HER2 FISH slides, the data was given as a ratio of the HER2 signal to the CEP 17 signal. Although we also included the average copy numbers of each signal, we did not specifically evaluate chromosome 17 polysomy or heterogeneous gene amplification as scores were only accepted if >40 tiles were counted. The discussion has been updated to address the discrepancy between IHC and FISH.

3. “Details of how the auto analysis system was “trained” (parameters used) are necessary, preferably with a visual example or graphic explanation.”

A representative core was chosen containing areas that would be scored as 1+ and 3+ visually. Using the color pickup tool within the Ariol image analyzer, we selected membranes with weak positive staining and assigned “1+ intensity” staining; we then selected the membranes with strong positivity and assigned “3+ intensity”. Similarly, we selected counterstained nuclei with the color pickup tool, and adjusted the desired size, roundness and other shape parameters under visual control. Numeric values for colors of the positive objects, i.e. membranes, and negative objects, i.e. nuclei, were stored on the hard drive in a color classifier file. Numeric values for the shape of the nuclei were stored in a separate shape classifier file. The program used these two files for segmentation of the nuclei and the membranes in all other cores, and these two files were sent out to be used in the machine 2. Scores from a “0” to a “3+” were automatically generated by the Ariol image analysis software for each core, based on the intensity and completeness of the positively stained membranes, and the percent of positive cells. The Ariol algorithm applies HercepTest criteria for the score calculations. Visual examples and a graphical explanation are given in Figure 1.

Minor essential revisions:
1. “Control for color / white balance is cited as a possible factor in the difference between the two machine results: the authors state “The descriptors of the color and shape of the positive and negative tumor cells were transferred from one system to another, therefore variations in the image analysis results depended only on the scanner settings, i.e. positioning and white balance, but not on the image analysis settings.” Therefore, the procedure for setting the white balance and positioning needs to be given in order to assess the extent to which these did vary and how this variation relates to the results.”

Calibration, slide positioning and white balance ensure accurate scanning and analysis. We performed brightfield calibration using the Calibration slide that comes with the system. The system was set to Kohler illumination to capture high quality images. The Methods section has been updated.

Discretionary Revisions:
1. “While not essential for the results, an explanation for the large differences between the 0 and 1+ scores between the visual inspection and the instrument scores would be of general interest since it points to an underlying difference between visual and digital analysis.”
Statistical analysis of data from only 1212 cases with all IHC data available shows that Ariol system does tend to score more 1+ cases in comparison to visual inspection. However, this would not change clinical approach as 0 and 1+ cases are both interpreted as negative. In practice, the accuracy of automated quantitative analysis depends on a variety of factors. In addition to technical factors, fully automated system ‘objectively’ scores any staining present in the core and cannot distinguish between malignant and benign lesions with a precision comparable to the expertise level of a pathologist. The discussion has been updated.

Language:

1. “There are some missing definite (“the”) and indefinite (“a”) articles, especially in the final two paragraphs of the discussion, e.g. “between experienced operator and pathologists,” and “with further refinement Ariol system could be used for clinical service.”

The paper has been checked for language.

Editorial comments

1) Copyediting- We recommend that you copyedit the paper to improve the style of written English.

The paper has been copyedited.

2) Competing interests- Please re-word your declaration of Competing interests according to our requirements. If there are none to declare, please write 'The authors declare that they have no competing interests'.

The declaration of Competing interests has been re-worded.

3) Conclusions- please include a conclusions section in your manuscript. Manuscript sections should include (in the following order): Abstract; Background; Methods; Results; Discussion; Conclusions; Abbreviations (if any); Competing interests; Authors' contributions; Acknowledgements; References; Figure legends (if any); Tables (if any); Description of Additional files (if any).

The conclusions section has been included.

4) Acknowledgements. “We strongly encourage you to include an Acknowledgements section between the Authors contributions section and Reference list. Please acknowledge anyone who contributed towards the study by making substantial contributions to conception, design, acquisition of data, or analysis and interpretation of data, or who was involved in drafting the manuscript or revising it critically for
important intellectual content, but who does not meet the criteria for authorship. Please also include their source(s) of funding. Please also acknowledge anyone who contributed materials essential for the study. Authors should obtain permission to acknowledge from all those mentioned in the Acknowledgements. Please list the source(s) of funding for the study, for each author, and for the manuscript preparation in the acknowledgements section. Authors must describe the role of the funding body, if any, in study design; in the collection, analysis, and interpretation of data; in the writing of the manuscript; and in the decision to submit the manuscript for publication.

The Acknowledgements section has been included.

5) We believe that your manuscript would be better suited to the scope of another of our journals, BMC Clinical Pathology (http://www.biomedcentral.com/bmcclinpathol/). Please let us know if you would be happy to have your manuscript transferred to BMC Clinical Pathology or if you would prefer for your manuscript to continue being considered in BMC Cancer. If you wish to transfer, we handle this on your behalf.

We would prefer for the manuscript to continue being considered in BMC Cancer.

6) Please also ensure that your revised manuscript conforms to the journal style (http://www.biomedcentral.com/info/ifora/medicine_journals). It is important that your files are correctly formatted.

The revised manuscript conforms to the journal style.

Sincerely yours,

Samuel Aparicio
Molecular Oncology and Breast Cancer Program
BC Cancer Research Centre
675 W 10th Avenue, V5Z 1L3, Vancouver, BC, Canada
Phone: 1 604 675 8200, Fax: 1 604 675 8218
Email: saparicio@bccrc.ca