Reviewer's report

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

Version: 1 Date: 10 December 2008

Reviewer: Emiel Janssen

Reviewer's report:

Major compulsory revisions

This is a well written article, but the statistics and number of patient in the different groups are very different from table to table and for the different calculations.

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

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.

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?

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?

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

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 agreement to be similar as well.
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?

Please note the publications by Skaland et all on Her2 digital assessment.

**Level of interest**: An article whose findings are important to those with closely related research interests

**Quality of written English**: Acceptable

**Statistical review**: Yes, and I have assessed the statistics in my report.

**Declaration of competing interests**: I declare no that I have no competing interest