Pathol. Oncol. Res. (2017) 23:79–84
DOI 10.1007/s12253-016-0091-4

ORIGINAL ARTICLE

Agreement of Different Methods for Tissue Based Detection of HER2 Signal in Invasive Breast Cancer

Gaurav Thakral[1] · Andrew Wey[1] · Mobeen Rahman[1] · Rui Fang[1] · Christopher Lum[1]

Received: 28 June 2015 / Accepted: 6 July 2016 / Published online: 14 July 2016

Abstract Breast cancer is the second leading cause of cancer mortality amongst American women. The HER2 gene encodes a cell surface receptor that affects cell proliferation and has been recognized as a diagnostic factor in treatment selection for invasive breast cancer. Examine accuracy in HER2 detection between manual count, computer assisted, and automated tiling algorithm. 42 randomly selected invasive breast cancer specimens were enumerated by fluorescence in situ hybridization (FISH)for HER2 and CEP17 markers using the Vysis HER2 assay (AbbotLaboratory, North Chicago, IL). Specimens were tested using three methods: Manual, computer assisted nuclei selection (Tissue FISH MetaSystems, Newton, MA), and automated enumeration (MetaSystems, Newton, MA). The greatest bias and widest agreement limits for HER2 and CEP17 were seen in Automatic versus Manual, the gold standard. HER2 values greater than 6 possessed the greatest bias and widest agreement limits. CEP17 comparison showed similar bias and agreement limits for each comparison. Kappa values indicated good agreement for all methods although Tissue FISH and Manual possessed better agreement. Higher agreement at lower HER2 & CEP17 count maybe due to fewer chromosomal aberrations, in which selection of field of views has less variation between methods. Alternatively, increased background signals seen in polyploidy may be responsible for the variations in signal count. Manual and Tissue FISH demonstrated good agreement amongst by both Altman Bland and Cohen's Kappa. While the automatic method has good agreement at lower HER2, the sharp increase in variability at higher HER2 counts illustrates a limitation of the automatic method.

Keywords Fish · Altman bland · HER2 Recptor · Breast neoplasms/diagnosis · Breast neoplasms/treatment

Introduction

Breast cancer is the second leading cause of cancer mortality amongst American women. Breast cancer, like all cancers, carries somatic mutations and chromosomal aberration in their genomes. A subset of genomic change, called driver mutations, give the mutated cell line a clonal selective advantage for oncogenesis [1, 2]. Screening for these mutations has advanced the diagnosis of breast cancer which in turn leads to more successful treatment and reduced mortality.

HER2 gene amplification has been recognized for many years as a diagnostic test and a critical factor in selection of treatment options. The HER2 cell surface receptor affects cell proliferation and survival of the cell line. The increased receptor expression serves as a target for the humanized monoclonal antibody Herceptin (Trastuzumab) which binds with high affinity to the extracellular domain of HER2 [3–6]. An international, multicentered study concluded treatment with Herceptin is well tolerated and prolongs life for those with an otherwise poor prognosis [4]. The ASCO (American Society of Clinical Oncology) and CAP (College of American Pathologist) guidelines mandate every invasive breast cancer case have HER2 enumeration to evaluate for treatment with Herceptin [3].

The FISH analysis utilizes 30 kilobase (kb) or larger probes and detects gene amplification and rearrangements. Manual scoring, the gold standard, involves a pathologist or technologist selecting fluorescent signals within 20 nuclei for enumeration which can be a lengthy and laborious process. A second method uses computer assisted touch screen monitor to

✉ Gaurav Thakral
gthakral@hawaii.edu

[1] University of Hawaii West, Honolulu, HI, USA

digitally select nuclei with a digital pen to encircle individual nuclei followed by a computerized enumeration. The user can review and adjust the signal count for minimization of background signals, exclusion of signals outside nuclei, and other enumeration adjustments. In the automated enumeration system, the software uses fixed squares to quantify signal counts. The size of each tile is generally the size of one or two nuclei with HER2 signals essentially counted over the entire tile. Tiles are randomly placed in select regions where fluorescent signals are highest and therefore can include tumor and non-tumor nuclei, background signals, and incomplete nuclei at the tile edge [7, 8].

Materials and Methods

Case Selection

A total of 41 invasive breast cancer cases were randomly selected from October 2013 to April 2014. Only two cases were metastatic cancer with primary breast, the remainder of specimens were invasive ductal carcinoma. Average age at diagnosis was 61.4 with a median of 60.5 years. All identified specimens underwent Automated and Tissue FISH at a tertiary cancer center.

Fish

HER2 copy number was evaluated using the Vysis FDA-approved HER2 DNA probe kit (Abbott Laboratory, North Chicago, IL). Deparaffinized 4 um tissue sections were immersed in 0.2 N HCL for 20 min, followed by pre-treatment solution at 98 °C for 30 min, and then subjected to protease digestion at 37 °C for 5 min. Slides were hybridized with HER2 DNA probe mixture containing HER2 DNA probes (labeled with Spectrum Orange) and CEP17 DNA probes (labeled with Spectrum Green). Glass coverslips were applied and then the slides were denatured at 74 °C for 2 min and hybridized overnight at 37 °C in a humidified hybridization chamber. Slides were then washed in a post-hybridization buffer at 73.5 °C for 2 min and dried in the dark. Nuclei were then counterstained with 10 uL of 4′,6-diamidino-2-phenylindole(DAPI). Slides were kept in the dark at 4 °C until signal enumeration.

Manual Screening

Manual screening was performed on Olympus fluorescent microscopes using oil immersion 60X and 100X objectives. After selection of 10 FOV, at least 50 nuclei were used for signal count. The mean HER2 signal and CEP17 signal spot counts were manually recorded and the HER2/CEP17 ratio was calculated.

Tissue FISH Selection Module

Tissue FISH screening was performed using a WACOM™ touch screen monitor. After selection of 10 FOV, at least 50 nuclei were digitally traced for signal count. Traced nuclei were enumerated for HER2 and CEP17 signal within the encircled area using MetaSystems. HER2/CEP17 ratios were calculated using the Tissue FISH capture algorithm.

Automated Screening

Automated screening was performed on the MetaSystem imager using a modified FDA cleared algorithm. At least 10 fields of view (FOV) were selected using an H&E reference. NucleiHER2 and CEP17 signals were counted and ratios calculated using a tile image capture algorithm.

Statistical Analysis

The AltmanBland method assesses the agreement between two measurement methods. In essence, two methods with high agreement can be used interchangeably. The x-axis shows the mean of the two methods ($[x^A + x^B]/2$), whereas the y-axis represents the difference between the two methods for each patient ($x^B - x^A$). The mean difference between two methods is the bias, while the 95 % agreement limits, which are the dashed lines in the figures, represent the interval that contains 95 % of the between-method differences. Thus, narrow 95 % agreement limits indicate good overall agreement, while wide agreement limits indicate poor agreement [9, 10]. The Altman Bland plots allow the investigation of systemic differences between methods, possible outliers, and relationship of discrepancies between measurements.

We did two separate Altman-Bland analyses. The first Altman-Bland analysis assesses overall agreement between the three different measurement methods for HER2, CEP17, and the HER2/CEP17 ratio. The second Altman-Bland analysis separately investigates the agreement of the three measurement methods in three distinct HER2 categories: 0–4, 4–6, and greater than 6, based on the ASCO/CAP guidelines. The separate categories allow the bias and the 95 % agreement limits to change as the underlying HER2 value increases. Finally, Cohen's Kappa assesses the between-method agreement for the final HER2 status. Cohen's kappa is a statistical measure of inter-rater agreement for categorical items and was used to measure the agreement that occurs beyond random chance. Values around 0.5 are considered to indicate moderate agreement, while values above 0.8 are considered to indicate good agreement [11].

Results

As demonstrated in Fig. 1, the greatest bias in the Altman Bland graphs for HER2 was seen in Automated versus Manual. In addition, both Automatic comparisons had wide agreement limits indicating high variability and poor agreement, while Manual versus Tissue FISH possessed the smallest bias and relatively narrow agreement limits, which indicate better agreement. For the analysis dividing HER2 into 0–4, 4–6, and greater than 6 subcategories, each comparison group exhibited a bias close to zero for the 0–4 and 4–6 HER2 categories with a large increase in bias for the HER2 greater than 6 category. This indicates that the Automatic method overestimates TissueFISH and Manual for HER2 values greater than 6. Lastly, for each comparison, the width of the 95 % agreement limits progressively increases with larger HER2 counts. Thus, the between-method variability increases with larger underlying HER2 values. The Automatic comparisons again possessed the widest agreements limits for each HER categorization, which indicates relatively poor agreement.

Figure 1 showed that the between-method agreement, in general, is better for CEP17 than HER2 as indicated by the relatively narrow agreement limits. The Automatic method overestimates the CEP 17 count compared to both Manual and TissueFISH as indicated by the positive bias. TissueFISH overestimates CEP17 count compared to Manual. Interestingly, there was no visual trend seen in the CEP17 plots, unlike the HER2 plots that show a trend at high values. Figure 2 presents the results of the Altman Bland analysis for the HER2/CEP17 ratio. Automatic versus Manual possesses the lowest variability for the HER2/CEP17 ratio. The variability of the HER2/CEP17 ratio increases for values greater than 4 for each comparison. However, the Automatic versus Manual possess a dispersed

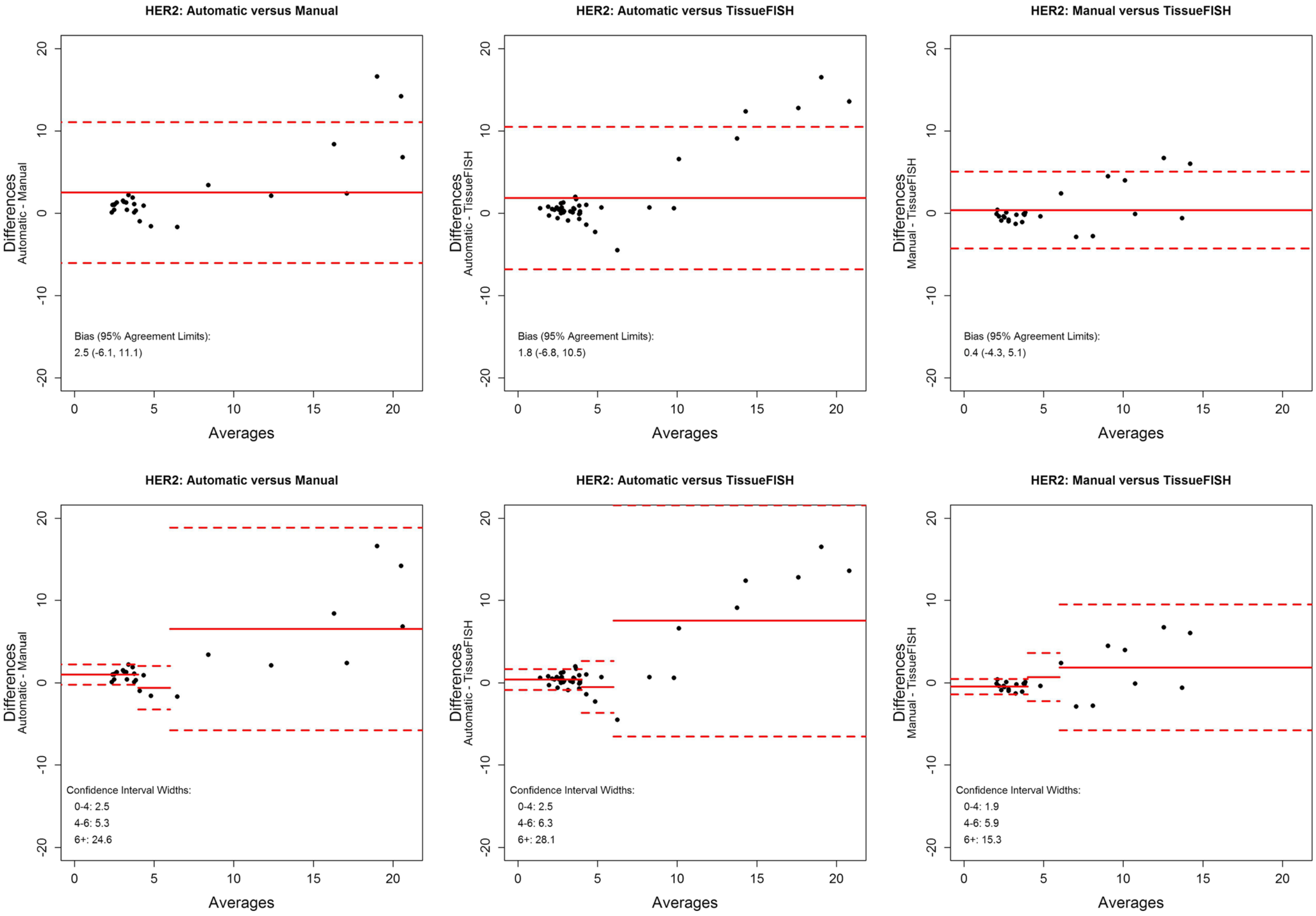

Fig. 1 Altman-Bland analysis for HER2. The first row is the overall analysis, while the second row separately investigates the agreement for the three distinct HER2 categories. Solid lines represent the bias, or the mean difference, between the two methods. Dashed lines are the 95 % agreement limits, which represent the interval that contains 95 % of the between-method differences

pattern, while the other two groups have a clustered pattern around the upper agreement limit.

Tables 1 and 2 present the bias and agreement limits for CEP17 and the HER2/CEP17 ratio, respectively, for each subcategory of HER2 values. For each method, the CEP17 bias increases with larger HER2 values. In addition, the CEP17 agreement limits, become progressively wider as the HER2 value increases although the Automatic versus Tissue FISH comparison is a notable exception at HER2 values greater than 6. Similarly, Table 2 depicts the bias and agreement limits for HER2/CEP17 ratio for each subcategory of HER2 values. For each comparison, the absolute bias increases slightly for the 4–6 range of HER2 values followed by a dramatic increase for the greater than 6 range of HER2 values. In addition, the agreement limits progressively widen as value of HER2 increases. Thus, both tables indicate that the between-method variability of CEP17 and HER2/CEP17 ratio increases as HER2 increases.

Table 3 presents the Kappa analysis. Automatic possesses moderate agreement with both TissueFish and Manual with Kappas slightly above 0.5. In contrast, Manual versus Tissue FISH had excellent agreement with a Kappa over 0.9.

Discussion

Herceptin (Trastuzumab) is the cornerstone treatment for women with HER2 gene amplified invasive breast cancer. Accurate assessment of HER2 and CEP17 is essential for identifying patients who may benefit from Trastuzumab therapy. The ASCO/CAP guidelines strongly encourage HER2 testing for patients with newly diagnosed invasive breast cancer or metastatic breast cancer [12]. While HER2 can be assessed by numerous methods, the clinical use of FISH is well established and widely used [12]. Our study demonstrates the variability of HER2 and CEP17 signals increasing with higher copy number. This variability may potentially result in misclassification of HER2 status, which is particularly a potential problem for the Automatic method.

In our study, HER2 signals showed an increase in between-method variability with increasing signals, while CEP17 did not possess higher variability with increased CEP17 signals although CEP17 variability did increase as the corresponding HER2 values increased. CEP17 was not stratified because the majority of values were in a narrow range of 2–4 signals. Considering that CEP17 agreement limits were similar for

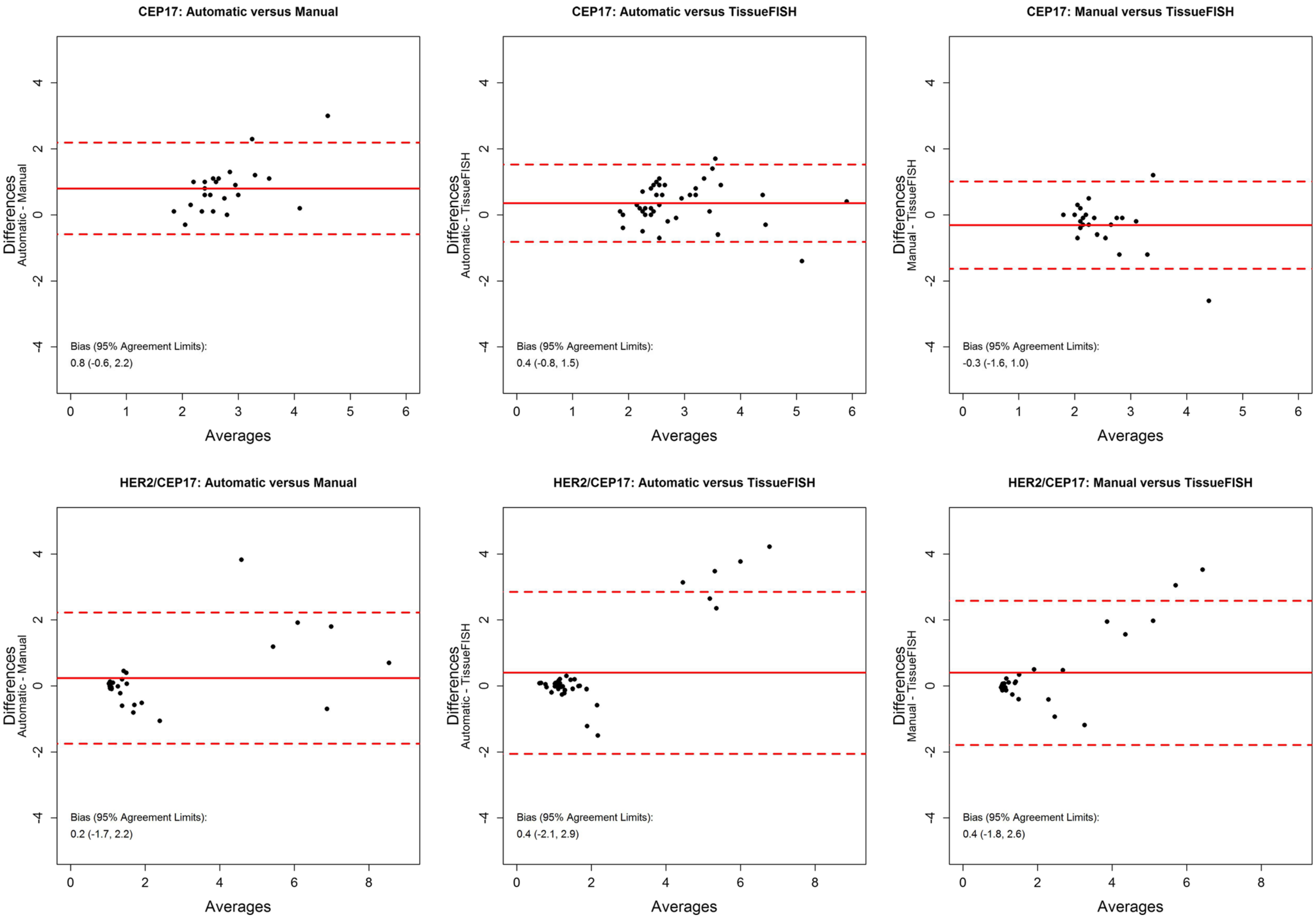

Fig. 2 Altman-Bland analysis for CEP17 (first row) and HER2/CEP17 (second row). Solid lines represent the bias, or the mean difference, between the two methods. Dashed lines are the 95 % agreement limits, which represent the interval that contains 95 % of the between-method differences

all three groups, one would expect the difference in HER2 agreement limits to be reflected in the HER2/CEP17 agreement limits. However this is not the case. Manual versus TissueFISH has a wider HER2/CEP17 agreement limits despite a relatively narrow HER2 agreement limits. The HER2/CEP17 variability is similar between the three methods despite the large difference in variability for HER2. Compared to other groups, Manual versus TissueFISH had the least variability for HER2, similar variability for CEP17, yet higher variability for HER2/CEP17. Thus, the variability of HER2/CEP17 was independent of HER2, which may be partly due to the differences in detection between Manual and Tissue FISH.

There was variation in the average HER2 and CEP17 count between Manual (13.9, 2.3 respectively), TissueFISH (9.9, 2.5 respectively), and Automatic (20.8, 2.9 respectively). Interestingly, the HER2 count for TissueFISH is less than half the count for Automatic, yet there is only a 9 % decrease in CEP17 signal detection. Since both methods utilize a computer algorithm for signal detection, a portion of HER2 signals are disregarded by the pathologist or technician. Alternatively, the Automatic method includes areas of tissue with extranuclear signals. Thus, as the lower agreement limit approaches a difference of zero for automatic with HER2 greater than 7, there is a near 95 % chance the automatic method will overestimate the single count or cellular tissue. This occurrence is likely due to a combination of tile sampling including more than one cell and inclusion of extranuclear signals by the automatic method [7, 8].

Lack of or minimal chromosomal aberration samples have a consistent distribution of HER2 and CEP17 signal throughout the tissue resulting in less variation between enumeration methods. Hence, there is less between-method variability seen at HER2 less than 5 and CEP17 less than 3. Alternatively increased background and extranuclear signals seen in polyploidy may be responsible for the variations in signal count. Tables 1 and 2 show the bias and variability in HER2/CEP17 and CEP17 marker increases as the average HER2 increases. The only exception was a decrease in variability of CEP17 for Automatic versus TissueFISH. At higher HER2 averages, there was more variability between all groups. As pathologist and technologist look at cellular areas, they are mindful to avoid overlapping cells or cells with boarders that cannot be deciphered. In contrast, the automatic method can not apply this criterion and, therefore, variability seen with automatic method is likely to occur from a consistent overestimation of signals, which is more pronounced in high HER2 amplification cases.

Table 1 The bias and 95 % agreement limits for CEP17 given a patient's HER2 categories

	Auto-Manual	Auto-FISH	Manual-FISH
HER2 Value 0–3	0.0 (−0.4, 0.5)	0.0 (−0.2, 0.3)	0.0 (−0.4, 0.3)
HER2 Value 2–6	−0.7 (−1.5, 0.2)	−0.5 (−1.5, 0.4)	0.0 (−0.8, 0.9)
HER2 Value 6+	1.0 (−2.1, 4.0)	2.0 (−2.0, 6.0)	1.3 (−2.1, 4.7)

Table 2 The bias and 95 % agreement limits for HER2/CEP17 ratio given a patient's HER2 categories

	Auto-Manual	Auto-FISH	Manual-FISH
HER2 Value 0–4	0.8 (−0.3, 2.0)	0.3 (−0.8, 1.4)	−0.4 (−1.0, 0.3)
HER2 Value 4–6	0.8 (0.0, 1.5)	0.4 (−2.0, 2.7)	0.2 (−0.4, 0.8)
HER2 Value 6+	0.8 (−1.2, 2.8)	0.4 (−0.4, 1.3)	−0.4 (−2.6, 1.8)

Kappa values indicate agreement between all methods, while Manual versus TissueFISH possess the best agreement. Similar to the Altman-Bland analysis, the Cohen's Kappa demonstrates that TissueFISH and Manual have the greatest degree of agreement for classifying final HER2 status. In particular, the Cohen's Kappa for the two automatic methods possess similar moderate levels of agreement, while Manual versus Tissue FISH shows near perfect Kappa and is considerably better than Automatic.

Altman Bland method assesses the extent to which two methods agree. It is often used to determine whether methods that one might replace the other with significant accuracy [10]. The overall HER2 Altman Bland plots have a small bias for each comparison group. In fact, Manual versus TissueFISH has an almost zero bias, indicating the mean HER2 from both methods are equal. Subcategorization of HER2 revealed a large bias and wide agreement limits at higher signal counts. The advantage of the subcategory approach is a visual illustration of the between-method agreement for different ranges of HER2 values. For example, in all three plots, the between-method variability progressively increases as the HER2 range increases. Thus, the between-method agreement becomes progressively worse as the HER2 value increases. This progressively worse agreement may be the cause behind the low Kappa values for the Automatic method although this deserves further investigation.

Our study demonstrates good agreement amongst TissueFISH and Manual, the gold standard, by both Altman Bland and Cohen's Kappa. While the automatic method showed a good bias and agreement at lower HER2 counts,

Table 3 The agreement on the final outcome between the three different methods as assessed by Cohen's Kappa. Values above .8 are considered good while values around.5 are considered moderate

Compared Methods	Kappa (95 % CI)
Automated vs. Manual	0.48 (0.23, 0.72)
Automated vs. TissueFISH	0.51 (0.29, 0.74)
Manual vs. TissueFISH	0.93 (0.79, 1.00)

the sharp increase in variability at higher count illustrates a limitation of this method. A larger study is need to measure variability of HER2 around 6 and increasing CEP17. The increase in variability of HER2/CEP17 and CEP17 with increasing HER2 subgroups maybe related to increased computer assisted signal detections.

In our study, the use of FISH with increasing HER2 and CEP17 copy number variability lead to between method variability in classification of HER2 status. The different methods tested in showed an increased variability as the signal count increased. Interestingly the variability of HER2/CEP17 was independent of the HER2 signal count. Furthermore the variation in HER2 count between methods was greater than that of CEP17. Our belief is the selection of cells by pathologist is instrumental in high signal count regions, as demonstrated by the Altman Bland and Cohen's Kappa analysis. The three methods showed the least variability at lower dispersed signal counts, however an increase and clustering of signals resulted in increased between method variability. Automatic enumeration maybe a cost effective approach for enumeration as long as borderline and high signal count cases are confirmed by a pathologist.

Financial support for this research comes in part from Everest Foundation and National Institute of Health grants: U54MD007548, P20GM103466, and G12MD007601. We would like to thank Bruce Mitchell and Nila Winn for their full support in data collections and laboratory support.

References

1. Humphrey LL, Helfand M, Chan BK, Woolf SH (2002) Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 137(5 Part 1): 347–360
2. Heim S, Mitelman F. (2011) Cancer cytogenetics: chromosomal and molecular genetic abberations of tumor cells: John Wiley & Sons
3. Wolff AC, Hammond ME, Hicks DG, Dowsett M, McShane LM, Allison KH, et al. (2013) Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. J Clin Oncol Off J Am Soc Clin Oncol 31(31):3997–4013. doi:10.1200/JCO.2013.50.9984
4. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, Smith I, et al. (2005) Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med 353(16): 1659–1672
5. Hanna WM, Ruschoff J, Bilous M, Coudry RA, Dowsett M, Osamura RY, et al. (2014) HER2 in situ hybridization in breast cancer: clinical implications of polysomy 17 and genetic heterogeneity. Mod Pathol: an official journal of the United States and Canadian Academy of Pathology, Inc. 27(1):4–18. doi:10.1038/modpathol.2013.103
6. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL (1987) Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 235(4785):177–182
7. Hicks DG, Tubbs RR. (2005) Assessment of the HER2 status in breast cancer by fluorescence in situ hybridization: a technical review with interpretive guidelines. Hum Pathol 36(3):250–261. doi:10.1016/j.humpath.2004.11.010.
8. Furrer D, Jacob S, Caron C, Sanschagrin F, Provencher L, Diorio C. (2013) Validation of a new classifier for the automated analysis of the human epidermal growth factor receptor 2 (HER2) gene amplification in breast cancer specimens. Diagn Pathol 8:17. doi:10.1186/1746-1596-8-17
9. Myles P, Cui J. I. (2007) Using the bland–Altman method to measure agreement with repeated measures. Br J Anaesth 99(3):309–311.
10. Bland JM, Altman D. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327(8476):307–310.
11. Viera AJ, Garrett JM. (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37(5):360–363.
12. Wolff AC, Hammond ME, Hicks DG, Dowsett M, McShane LM, Allison KH, et al. (2014) Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. Arch Pathol Lab Med 138(2):241–256. doi:10.5858/arpa.2013-0953-SA