Interest of Artificial Intelligence Algorithms to Determine HER2 Low Status in Breast Cancer

Thomas Vidal[1], Kim Vianey[1], Cécile Maisin[1], Cécile Hayem[2], Gilles Benaim[1], Emilie Courcet[1], Olivier Deroo[1], Nicolas Hamant[1], Anne Le Hemon-Lepaul[1], Anthony Jacquier[1,3], Jean-Pierre Machayekhi[1], Nelly Youssef-Provençal[1,3], Pierre Serra[1], Charlène Vigouroux[1,3], Ossama Yacoub[1], Magali Lacroix-Triki[4], Marie Brevet[1]

1. Cypath & Cypath-RB, Villeurbanne, France ; 2 ACEFAS, Versailles, France ; 3 Dermapath, Villeurbanne, France ; 4 Gustave Roussy Cancer Campus, Villejuif, France

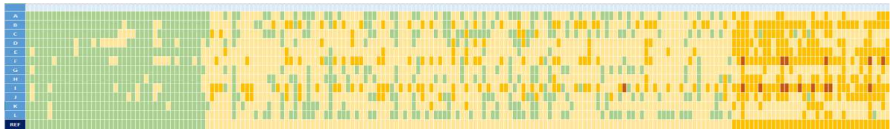
Background : Novel anti-HER2 antibody drug conjugates (ADCs) have shown efficacy in invasive breast cancers (BCs) expressing low levels of HER2 (1+ by immunohistochemistry (IHC) and 2+ non-amplified) (1). However, differentiating scores 0 from 1+ is challenging even for experienced pathologists (2). In private practice, pathologists frequently diagnose BCs and must be able to establish HER2 low status with reliability and reproducibility. Artificial intelligence (AI) algorithms based on deep learning could help in assessing HER2 low score. The objectives of this study are 1) to assess the concordance of HER2 low diagnosis amongst pathologists practicing in a liberal structure and 2) to study AI algorithm performance in this indication.

Design: 200 centrally stained HER2 IHC slides from primary BCs expressing low levels of HER2 were selected based on pathology reports. The slides were digitized and whole slide images (WSI) were then subjected to an inter-observer study including 12 pathologists from a multi-site private laboratory. The same 200 WSI were analyzed by a pathologist expert in breast pathology and then subjected to four AI tools designed for HER2 scoring. To study AI performance, AI tests were performed on regions of interest (ROI) representative of HER2 labeling. In parallel, the ease of use of each software and algorithm was evaluated.

This study was conducted in compliance with French ethical rules. The authors thank AstraZeneca for financial support and Mindpeak, Roche, Tribun Health & Visiopharm for providing the AI tests. MLT declares honoraria and institutional grants from AstraZeneca and Daïchi Sankyo. MB declares honoraria from AstraZeneca and Tribun Health.

Results: After proofreading by the expert pathologist, the cohort included 41 “IHC 0" cases, 120 “IHC 1+" cases and 36 “IHC 2+" cases by IHC (n=197 ; 3 slides excluded). With the expert pathologist as reference, the inter-observer studies showed an average “scoring accuracy” (cases 0+ versus 1+ versus 2+ versus 3+) of 68.4% (range: 54.8-87.8%) and an average “clinical accuracy” (cases 0+ versus group 1+/2+/3+) of 83.6% (range : 69.0-96.4%) (Figure 1).

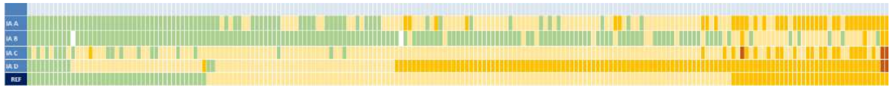
Figure 1: Evaluation of HER2 scores: pathologists vs. gold standard



🟩 HER2 0 ; 🟨 HER2 1+ ; 🟧 HER2 2+ ; 🟥 HER2 3+ ; Reference score is indicated at the bottom of the figure.

Figure 2 and table 1 present the accuracy for each AI tools regarding the “scoring accuracy” and the “clinical accuracy”. Sensitivity and specificity evaluate the algorithm ability to detect a true positive or true negative staining (“0+” vs “1+/2+/3+”) while “balanced clinical accuracy” is the mean between sensitivity and specificity. For each AI tool, pathologists gave a score concerning the viewer (fluidity, ease of use), the algorithm (how to launch it, how to see the result…) and the report generated by the tool. One AI tool was very easy to use while the other algorithms were more time consuming with a less user-friendly interface.

Figure 2 : Evaluation of HER2 scores: AI vs. gold standard



🟩 HER2 0 ; 🟨 HER2 1+ ; 🟧 HER2 2+ ; 🟥 HER2 3+ ; Reference score is indicated at the bottom of the figure.

Table 1 : A) Scoring accuracy of AI tools ; B) Clinical accuracy of AI tools

TEST	Scoring accuracy (%)
AI A	73.1
AI B	28.2
AI C	75.1
AI D	43.1

TEST	Unbalanced clinical accuracy (%)	Sensitivity (%)	Specificity (%)	Balanced clinical accuracy (%)
AI A	81.2	76.3	100	88.1
AI B	41.5	26.5	100	63.2
AI C	85.3	98.1	36.6	67.3
AI D	83.2	98.7	24.4	61.6

Table 2 : Assessment of AI tools

TEST	Viewer /5	Algorithm /5	Report /5	Global evaluation /5
AI A	4,7	4,6	4,4	4,6
AI B	3,3	3,6	3,7	3,3
AI C	2,8	2,8	2,9	2,8
AI D	4,2	4,3	3,9	4,1

Discussion: Inter-observer agreement for HER2 scoring within a liberal group is similar to that observed in the literature (2,3,4,5). Training for accurately establishing low HER2 status, which is currently ongoing all over France, should increase concordance between observers. AI tool may allow good reproducibility between pathologists with different training.

Among the 4 tools tested, results varied in terms of accuracy, sensitivity and specificity. Several factors may explain these differences: guidelines used to configure the HER2 scoring, their ability to recognize artefacts and lack of algorithm calibration with regard to Cypath slides... Moreover, these algorithms were not trained specifically for the HER2 low category and a new specific training for HER2 low will probably improve the performance for some of them.

Among the 4 tools, the AI A was clearly preferred by pathologists in terms of ease of use of the platform and speed to obtain the HER2 score result.

Conclusions: Because of accuracy variability, use of these AI tools in routine practice will require prospective validations. It is important to determine the exact question the algorithm has to answer.

Ease of use and integration of an algorithm will determine the final choice, at equivalent performance rate.

References: 1) S Modi et al. NEJM 2022 2) Fernandez AI, JAMA Oncol 2022 3) F Schettini et al. Breast Cancer 2021 ; 4) K Lambein et al. AJCP 2013