Agreement Across 10 Artificial Intelligence Models in Assessing HER2 in Breast Cancer Whole Slide Images: Findings from the Friends of Cancer Research Digital PATH Project

Brittany McKelvey[1], Pedro A. Torres-Saavedra[2], Jessica Li[2], Glenn Broeckx[3], Frederik Deman[3], Siraj Ali[4], Hillary Andrews[1], Salim Arslan[5], Santhosh Balasubramanian[6], J. Carl Barrett[7], Peter Caie[8], Ming Chen[9], Daniel Cohen[10], Tathagata Dasgupta[11], Brandon Gallas[12], George Green[13], Mark Gustavson[14], Sarah Hersey[15], Ana Hidalgo-Sastre[14], Shahanawaz Jiwani[16], Wonkyung Jung[4], Kimary Kulig[17], Vladimir Kushnarev[18], Xiaoxian Li[19], Meredith Lodge[8], Joan Mancuso[20], Mike Montalto[21], Satabhisa Mukhopadhyay[11], Matthew Oberley[9], Pahini Pandya[5], Oscar Puig[22], Edward Richardson[23], Alexander Sarachakov[18], Or Shaked[22], Mark Stewart[1], Lisa M. McShane[2], Roberto Salgado[3], Jeff Allen[1]

[1]Friends of Cancer Research, [2]Division of Cancer Treatment and Diagnosis, National Cancer Institute, [3]ZAS Hospitals, [4]Lunit, [5]Panakeia, [6]PathAI, [7]Univeristy of North Carolina at Chapel Hill, [8]Indica Labs, [9]Caris Life Sciences, [10]GlaxoSmithKline, [11]4D Path, [12]Center for Devices and Radiological Health, U.S. Food and Drug Administration, [13]GA Green Consulting LLC, [14]AstraZeneca, [15]Bristol Myers Squibb, [16]Molecular Characterization Laboratory, Frederick National Lab, National Cancer Institute, [17]Kulig Consulting, [18]BostonGene, [19]Emory University, [20]Patient Advocate, [21]Amgen, [22]Nucleai, [23]Merck & Co., Inc.

Abstract P2-02-18
Corresponding author:
bmckelvey@focr.org

Introduction

- Recent successes of HER2 antibody-drug conjugates (ADCs) have expanded patient eligibility for HER2-targeted therapy.
  - Accurate and consistent identification of patients who may benefit from ADCs, by assessing HER2 expression, is critical.
- Previous studies of agreement in HER2 scoring between pathologists highlight areas of discordance.[1]
- AI models have the potential to deliver more quantitative and reproducible HER2 assessments.
  - Large-scale comparative evaluations of these models' performance are currently lacking.
- Friends of Cancer Research created a research partnership, The Digital PATH Project, to describe and evaluate the agreement of HER2 assessment across independently developed AI models.

Materials & Methods

Patient Samples

Whole slide images (WSIs), both H&E-stained and HER2 IHC (N=1,124), from patients diagnosed with breast cancer in 2021 (N=733) were obtained from a single laboratory (ZAS Hospital, Antwerp, Belgium). Available pathology and specimen metadata include HER2 (ASCO/CAP[3]) scoring by three breast pathologists and information on slide processing and digitization.

Computational Pathology Models

Known commercial developers of HER2 computational pathology models were invited to participate in the project, resulting in 9 developers representing 10 models. Model attributes (e.g., input WSI type, HER2 output, key training/validation methods) were provided by the developers. The 10 AI models assessed HER2 status on all cases.

Statistical Analysis

A defined reference standard was not used. Agreement was evaluated using the overall percent agreement (OPA) and Cohen's kappa coefficient for all possible pairings of models across samples. Statisticians from the NCI Biometric Research Program performed independent analyses of pairwise comparisons of each models' HER2 outputs to determine the level of agreement. Results shown evaluate agreement across the 7 models providing predicted ASCO/CAP scores.

Model Characteristics

WSI* Model Inputs	# Models
HER2 IHC** only	7
H&E only	2
Both HER2 IHC and H&E	1

*Aperio GT 450 scanner, **anti-HER2/neu (4B5) Rabbit Monoclonal Primary

Model Outputs	# Models
Predicted ASCO/CAP Scores[3] (0, 1+, 2+, 3+)	**7**
Percent Tumor Cells Exhibiting Staining HER2 0, 1+, 2+, 3+	5
H-scores[4] (0-300)	6
HER2 Positive vs. Negative (as defined by the developer)	3
HER2 Negative vs. Low vs. Positive (as defined by the developer)	3
HER2 Status Includes Ultra-Low (as defined by the developer)	2
Continuous Scores (Not Binned)	4
Continuous Scores with Bins of HER2 0, 1+, 2+, 3+	4
Area of Invasive Carcinoma	7
Number of Invasive Carcinoma Cells	5
Number of Stained Invasive Carcinoma Cells	4

The 7 models providing predicted ASCO/CAP scores included 6 using HER2 IHC WSIs only and 1 using both HER2 IHC and H&E WSIs.

Results

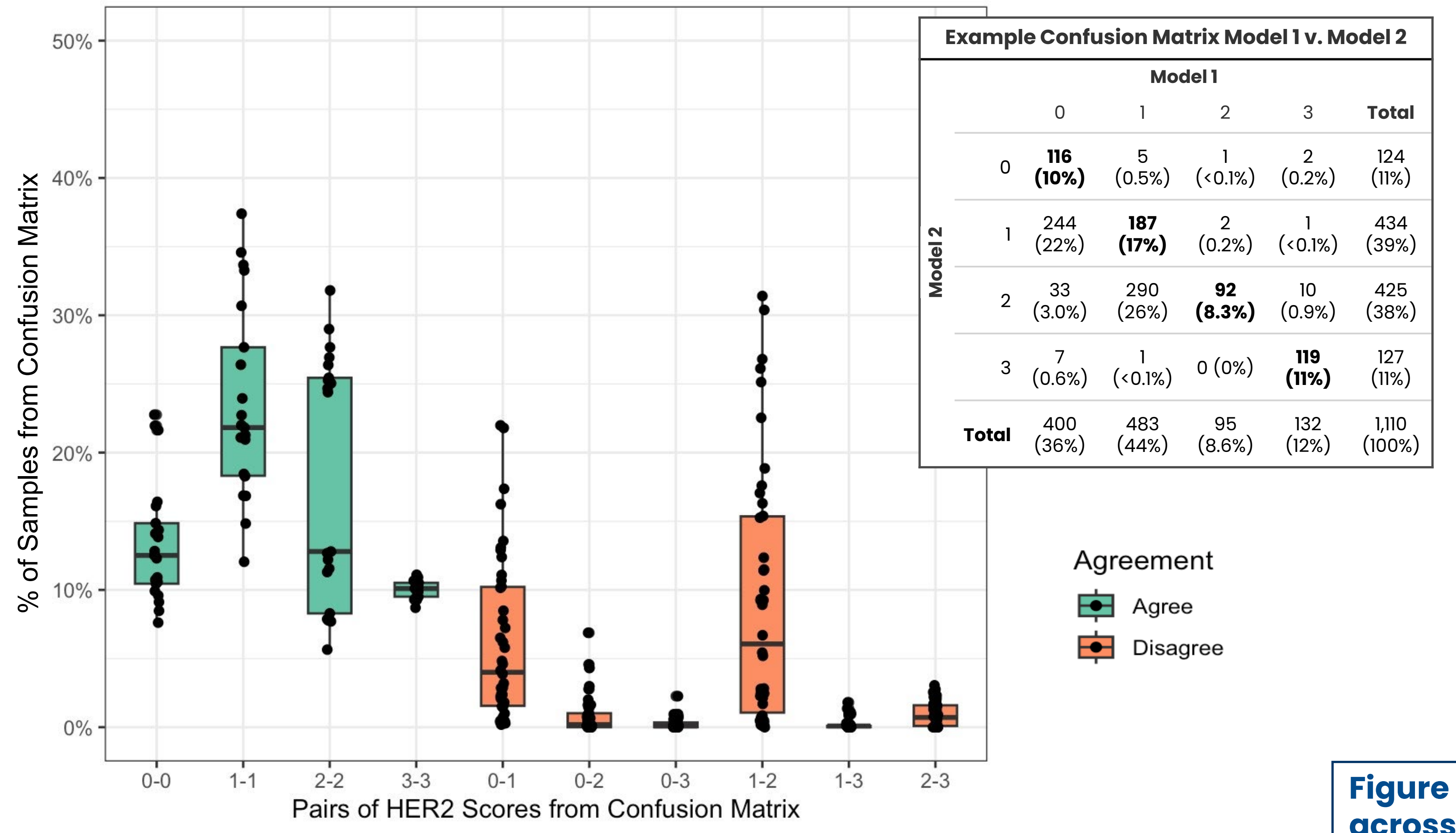
	n (%)
Age at Sample Collection	Median: 65
< 50	208 (18.5%)
50-64	336 (29.9%)
65+	580 (51.6%)
Diagnosis (Dx) History	
De Novo Dx of Breast Cancer	1060 (94.3%)
Recurrence	64 (5.7%)
Sex	
Male	16 (1.4%)
Female	1108 (98.6%)
Histological Grade	
1	149 (13.3%)
2	702 (62.5%)
3	231 (20.6%)
Not Recorded	42 (3.7%)
Histology	
Ductal	879 (78.2%)
Lobular	172 (15.3%)
Mucinous	25 (2.2%)
Other	48 (4.3%)

	n (%)
Clinical Stage	
I	612 (54.4%)
II	363 (32.3%)
III	85 (7.6%)
IV	64 (5.7%)
ER Status	
Positive	963 (85.7%)
Weakly Positive	15 (1.3%)
Negative	146 (13.0%)
PR Status	
Positive	815 (72.5%)
Negative	309 (27.5%)
Ki-67 Status	
0-10%	537 (47.8%)
11-100%	523 (46.5%)
Unknown	64 (2.7%)

Table 1: Clinical and demographic characteristics. Sample set clinical characteristics largely reflect the broader population of patients diagnosed with breast cancer.[2] Race and ethnicity data are unavailable (not captured at the site).

Agreement Measure Median (IQR)	**Categorical** **(0, 1+, 2+, 3+)**	**Binary** **(0 vs. 1+, 2+, 3+)**	**Binary** **(0, 1+ vs. 2+, 3+)**	**Binary** **(0, 1+, 2+ vs. 3+)**
OPA	65.1 (60.3, 69.06)	85.6 (82.43, 87.97)	79.9 (72.02, 82.18)	97.3 (95.90, 97.91)
Cohen's kappa	0.51 (0.45, 0.55)	0.57 (0.51, 0.61)	0.59 (0.44, 0.65)	0.86 (0.82, 0.90)

Table 2: Pairwise Agreement for Categorical and Binary Predicted ASCO/CAP HER2 Scores. There is high agreement in assigning HER2 3+ across the seven models, as well as assigning HER2 0, with larger variability in assigning 1+, 2+. Median and interquartile range (IQR) reported.



Example Confusion Matrix Model 1 v. Model 2

		Model 1				
		0	1	2	3	**Total**
Model 2	0	**116** **(10%)**	5 (0.5%)	1 (<0.1%)	2 (0.2%)	124 (11%)
	1	244 (22%)	**187** **(17%)**	2 (0.2%)	1 (<0.1%)	434 (39%)
	2	33 (3.0%)	290 (26%)	**92** **(8.3%)**	10 (0.9%)	425 (38%)
	3	7 (0.6%)	1 (<0.1%)	0 (0%)	**119** **(11%)**	127 (11%)
	Total	400 (36%)	483 (44%)	95 (8.6%)	132 (12%)	1,110 (100%)

Figure 2: Models disagreed most often on calling a sample 1+ or 2+, followed by 0 or 1+. There are fewer samples where model pairs assigned samples as two or more HER2 scores apart (e.g., 0-2, 1-3, 0-3). Dots represent the percentages from the confusion matrices of all 21 distinct pairs of models providing predicted HER2 scores (example confusion matrix to the right). Agree indicates model pairs assigned the same HER2 scores. Disagree indicates model pairs assigned different HER2 scores.

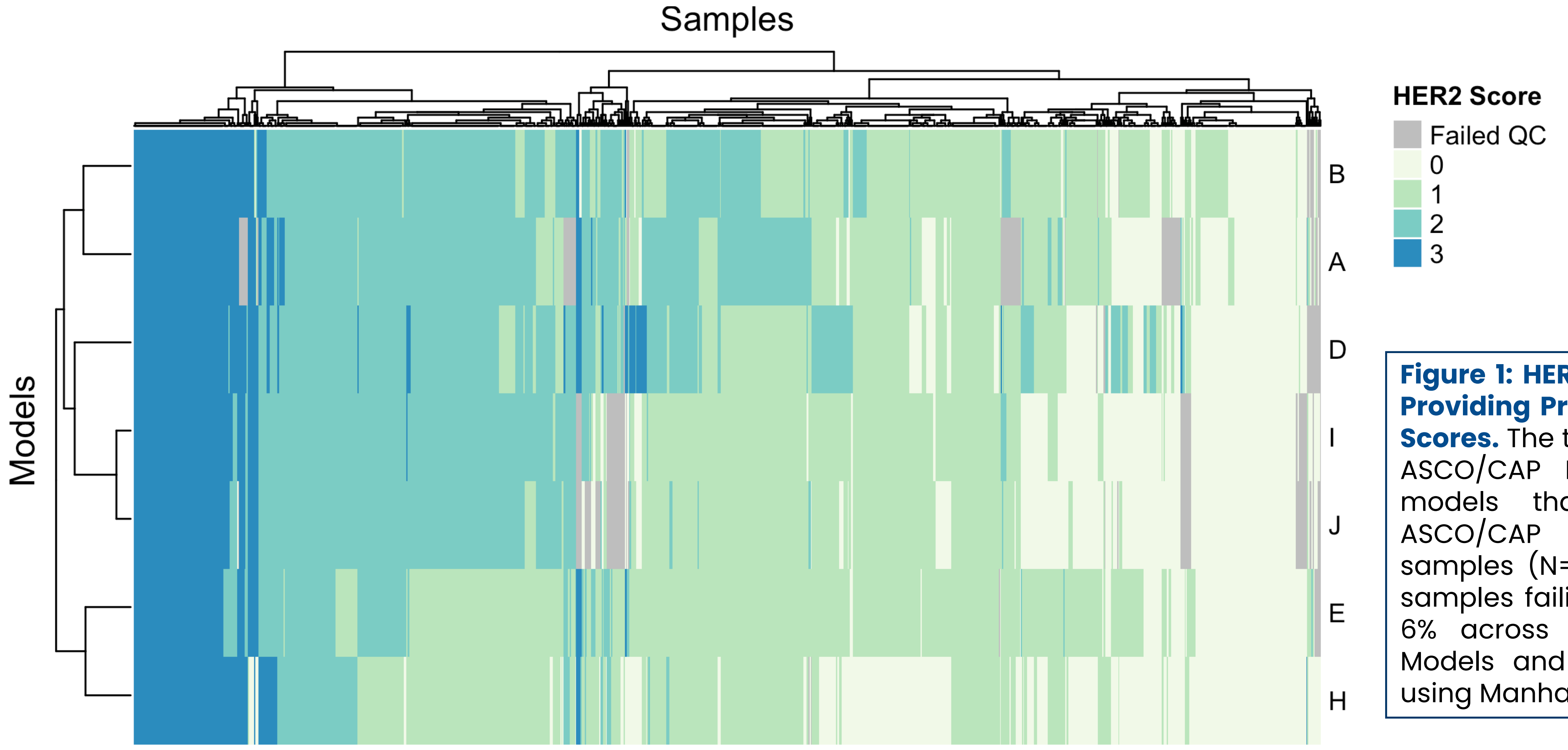


Figure 1: HER2 Scores Across Models Providing Predicted ASCO/CAP HER2 Scores. The tile plot depicts predicted ASCO/CAP HER2 scores by the 7 models that provided predicted ASCO/CAP HER2 scores for all samples (N=1,124). The proportion of samples failing QC ranged from 0.2-6% across models (median 1.5%). Models and samples are clustered using Manhattan distance.

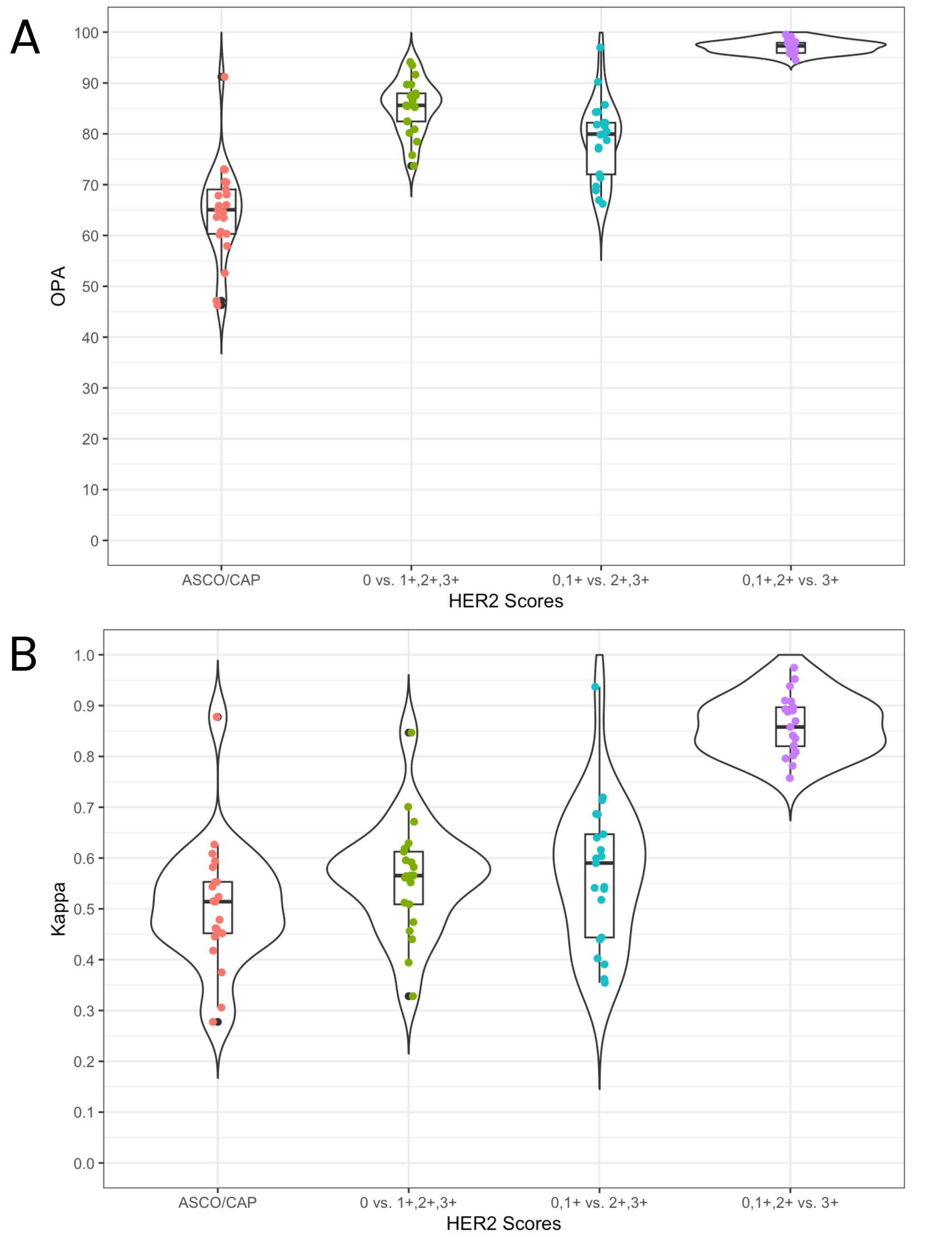


Figure 3: There is modest agreement in assigning HER2 scores across models (A: OPA, B: kappa). The highest overall agreement across the seven models providing predicted ASCO/CAP HER2 scores is in assigning samples as 3+ versus not 3+. Models also have higher OPA in assigning samples as 0 vs. not 0 compared to 0 or 1+ versus 2+ or 3+. Dots represent agreement measures for each model pair (21 pairs). "ASCO/CAP" represents categorical, i.e., 0 vs. 1+ vs. 2+ vs. 3+.

Conclusions

This unique partnership allowed us to assess the agreement of HER2 biomarker assessment across computational pathology models developed independently.

- Cases reported as HER2 3+ had the least variability and highest level of agreement across models.
- Cases reported as HER2 1+ and 2+ had larger inter-model variations observed.
- The trends in level of agreement between models across HER2 reported scores is similar to published agreement measures between pathologists.[1]

This ongoing partnership will enable a greater understanding of the variability across AI models under development and support establishing best practices for measuring and reporting AI-driven biomarker assessments in drug development and clinical practice, as well as informing approaches for the use of reference sets.

Next Steps

- Additional analyses on this cohort are underway, informing:
  - The association between the level of agreement of ASCO/CAP HER2 scores with patient, specimen, and model characteristics.
  - The agreement between models providing predicted ASCO/CAP HER2 scores and pathologist readings.
  - The level of agreement of H-scores between models.
  - The level of agreement between models that report the percentage of cells exhibiting the staining patterns associated with each ASCO/CAP category.
  - The level of agreement of HER2 scores between models that report other HER2 categories/scores.
- Friends will host a **Public Meeting on February 4th** to report findings, discuss policy implications, and provide recommendations for developing reference sets.

References

1: CJ Robbins, *et al.* Mod Pathol. 2023 PMID: 36788069.; 2: Surveillance, Epidemiology, and End Results Program, 2021. NCI, DCCPS.; 3: AC Wolff, *et al.* J Clin Oncol. 2018 PMID: 29846122. 4: K Jensen, et al. Mod Pathol. 2017 PMID: 27767098