FRIENDS of CANCER RESEARCH

Digital and Computational Pathology Tool Harmonization (PATH) Project

Mark Stewart, PhD

On behalf of the Digital PATH Project Working Group

Why HER2 and AI Matter

- HER2 is a clinically relevant biomarker in breast cancer, guiding treatment decisions
- Emerging therapies (e.g., antibody-drug conjugates) targeting HER2 are also effective in patients with “low” and “ultra-low” HER2 expression, expanding the eligible patient population and making precise, reproducible HER2 scoring increasingly important
- AI tools may help address challenges in reproducibility, accuracy, and scalability in HER2 scoring

FRIENDS of CANCER RESEARCH

The Role of AI Tools in HER2 Scoring

Reproducibility
AI models could help reduce inter- and intra-observer variability compared to manual scoring

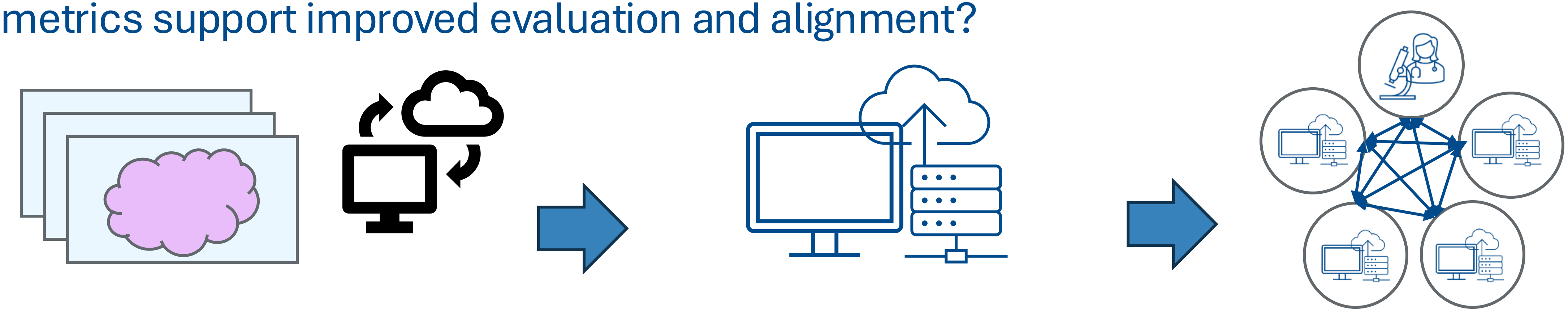
Efficiency
High-throughput capabilities for analyzing and sorting large datasets

Granularity
Potential for more nuanced analyses that may be infeasible for a human eye

Clinical implication
Accurate HER2 scoring is increasingly important with the expansion of therapies for patients with HER2-low

Digital PATH Project Approach

The Research Question: What factors contribute to variability in biomarker assessment across computational pathology platforms and what performance metrics support improved evaluation and alignment?



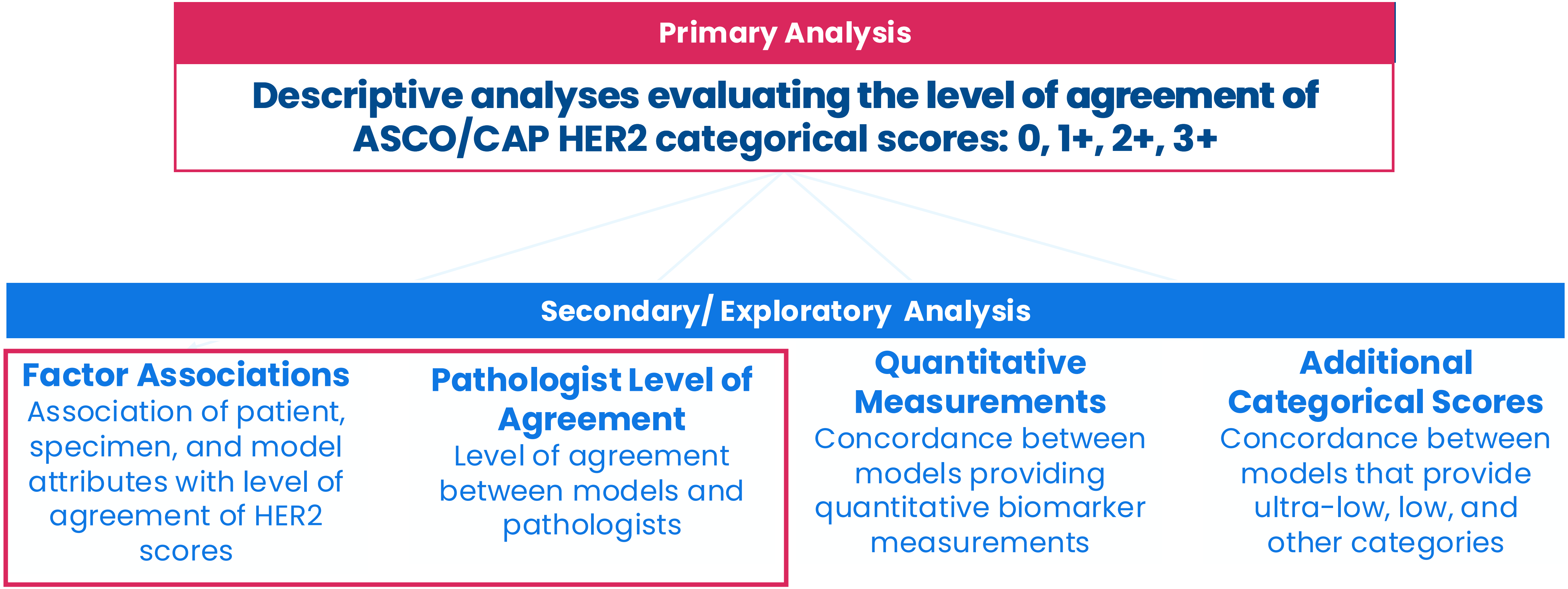
Develop a common dataset of >1000 breast cancer WSIs (digital images of HER2 IHC and H&E slides) and share with tool developers

Tool developers apply independently developed AI models to assess HER2 scoring

Compare results with pathologists from a single institution and among models to evaluate variability

The structured approach enables a systematic evaluation of variability and sources of discordance

Analysis Strategy Overview



Primary Analysis

Descriptive analyses evaluating the level of agreement of ASCO/CAP HER2 categorical scores: 0, 1+, 2+, 3+

Secondary/ Exploratory Analysis

Factor Associations
Association of patient, specimen, and model attributes with level of agreement of HER2 scores

Pathologist Level of Agreement
Level of agreement between models and pathologists

Quantitative Measurements
Concordance between models providing quantitative biomarker measurements

Additional Categorical Scores
Concordance between models that provide ultra-low, low, and other categories

Sample and Specimen Characteristics

Clinical/Tumor Characteristics

Histological Grade	n (%)
1	149 (13%)
2	702 (62%)
3	231 (21%)
Not Recorded	42 (4%)
Histology	
Ductal	879 (78%)
Lobular	172 (15%)
Mucinous	25 (2%)
Other	48 (4%)
Clinical Stage	
I	612 (54%)
II	363 (32%)
III	85 (8%)
IV	64 (6%)
ER Status	
Positive	963 (86%)
Weakly Positive	15 (1%)
Negative	146 (13%)
PR Status	
Positive	815 (73%)
Negative	309 (28%)
Ki-67 Status	
0-10%	537 (48%)
11-100%	523 (46%)
Unknown	64 (6%)

Demographics

	n (%)
Age at Sample Collection (yrs)	Median: 65
< 50	208 (19%)
50-64	336 (30%)
65+	580 (52%)
Diagnosis History	
De Novo Dx of Breast Cancer	1060 (94%)
Recurrence	64 (6%)
Sex	
Male	16 (1%)
Female	1108 (99%)

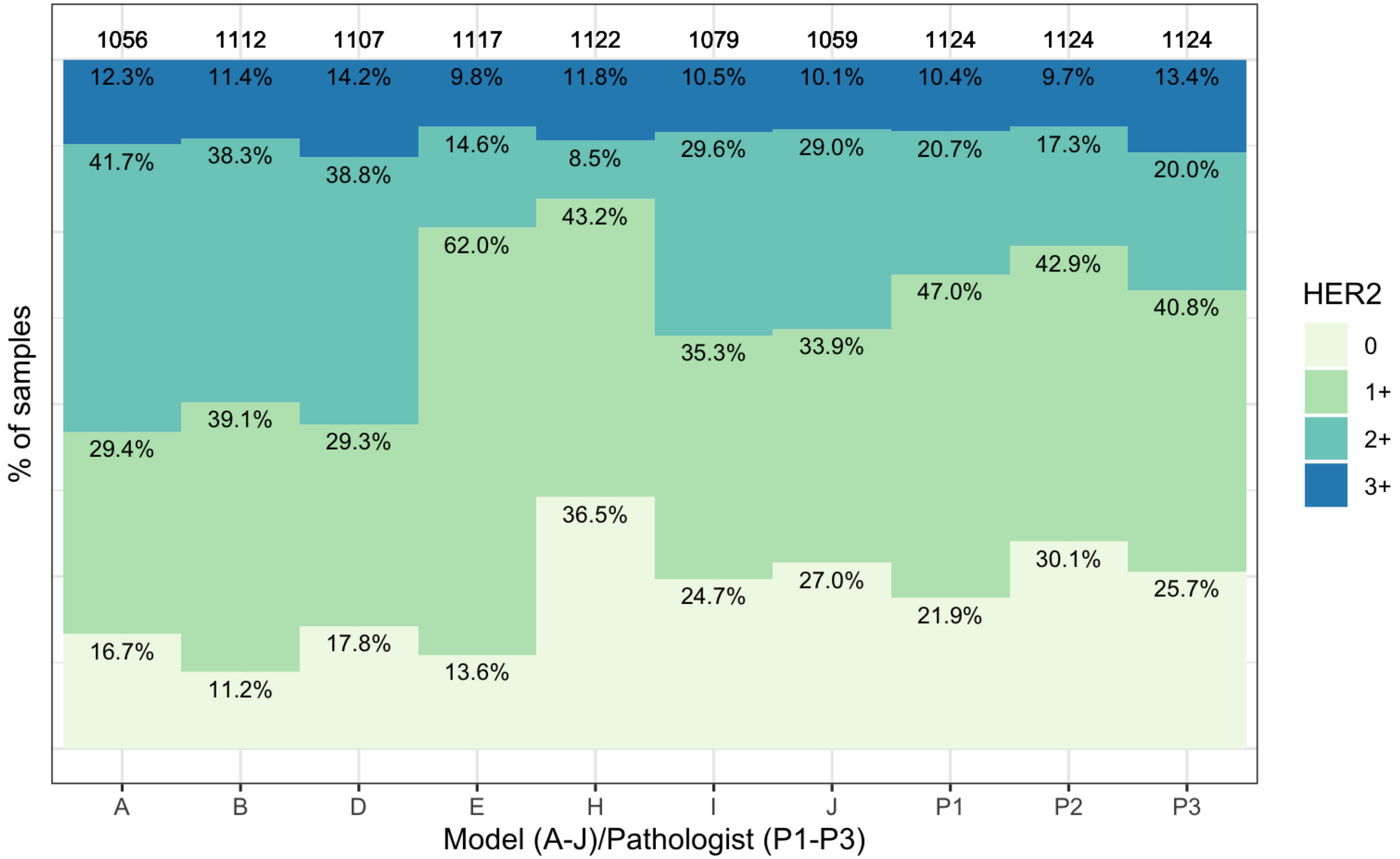
Clinical/ tumor characteristics and demographics align closely with population-level sources

Specimen Characteristics

Parameter	Details/Specification
Thickness of Tissue Section	4 micron
Fixation Type	Formaldehyde 4%
Coverslip	Sakura TissueTek Film
Hematoxylin Type	Hematoxylin II counterstain
Hematoxylin Time	12 minutes
Fixation Temperature	Room temperature
Mounting Media	Xylene
HER2 Antibody Clone	4B5
HER2 Antibody Manufacturer	Roche Ventura
Scanner Type	Leica Aperio GT 450 DX
Scanning Magnification	40x
Scanning Software Version	4.4

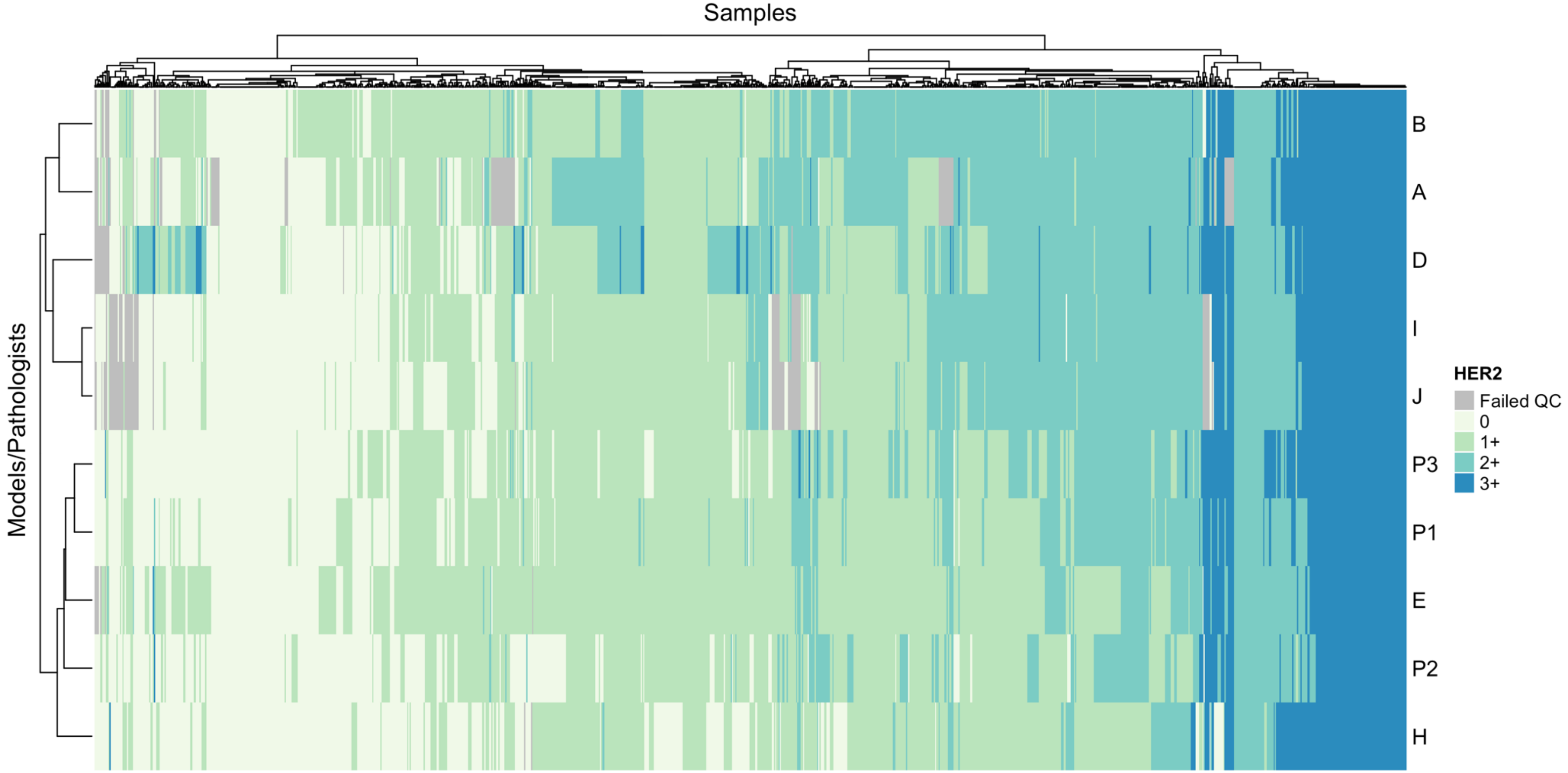
The specimen characteristics are homogenous to determine sources of variability in the model outputs.

HER2 Scoring Distribution Across AI Models and Pathologists



Finding: There is variability in HER2 outputs across models/pathologists, with more variability in 1+ and 2+ calls compared to 3+ calls.

HER2 Scoring Variability Across AI Models and Pathologists



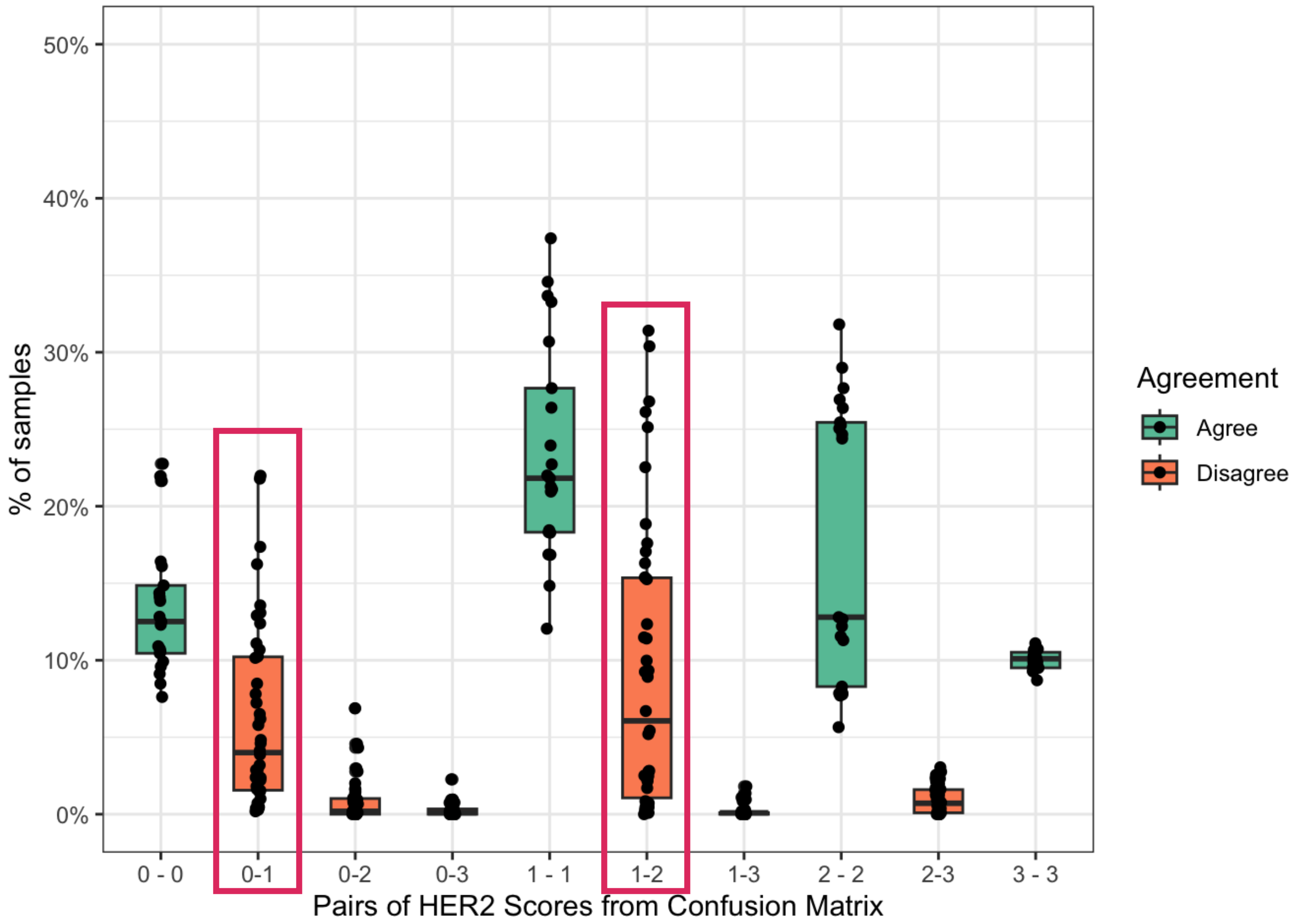
Finding: HER2 outputs show variability across AI models and pathologists, with the highest variability observed in 1+ and 2+ scores, while 3+ scores demonstrate greater consistency.

HER2 Scoring Agreement Across AI Models and Pathologists

	# of pairwise comparisons	Agreement Measure, Median	Categorical (ASCO/CAP)	Binary		
			(0, 1+, 2+, 3+)	(0 vs. 1+, 2+, 3+)	(0, 1+ vs. 2+, 3+)	(0, 1+, 2+ vs. 3+)
Models Only (7)	21	OPA (%)	65.1	85.6	79.9	97.3
Models (7) vs. Pathologists (3)	21		65.1	84.6	81.1	96.7
Pathologists Only (3)	3		70.4	85.1	86.3	96.6
Models Only (7)	21	Kappa	0.51	0.57	0.59	0.86
Models (7) vs. Pathologists (3)	21		0.51	0.57	0.58	0.84
Pathologists Only (3)	3		0.57	0.61	0.67	0.84

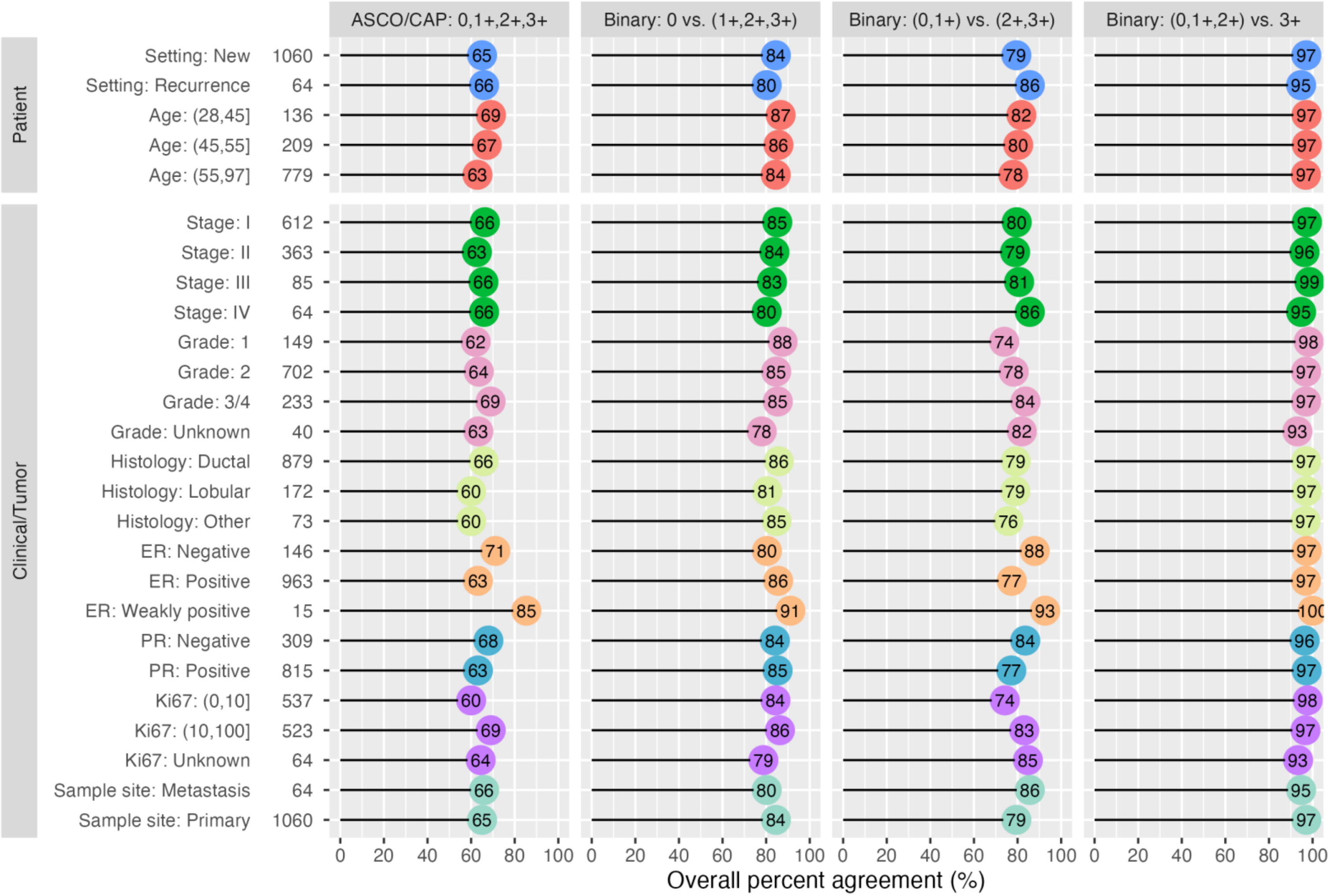
Finding: AI models and pathologists show similar HER2 scoring agreement, with the highest concordance for 3+ cases.

Pairwise Agreement in HER2 Scoring Across Models



Finding: Disagreements were more frequent between adjacent HER2 scores (e.g., 0 vs. 1+ or 1+ vs. 2+) rather than between more distant scores (e.g., 0 vs. 2+, 0 vs. 3+, or 2+ vs. 3+).

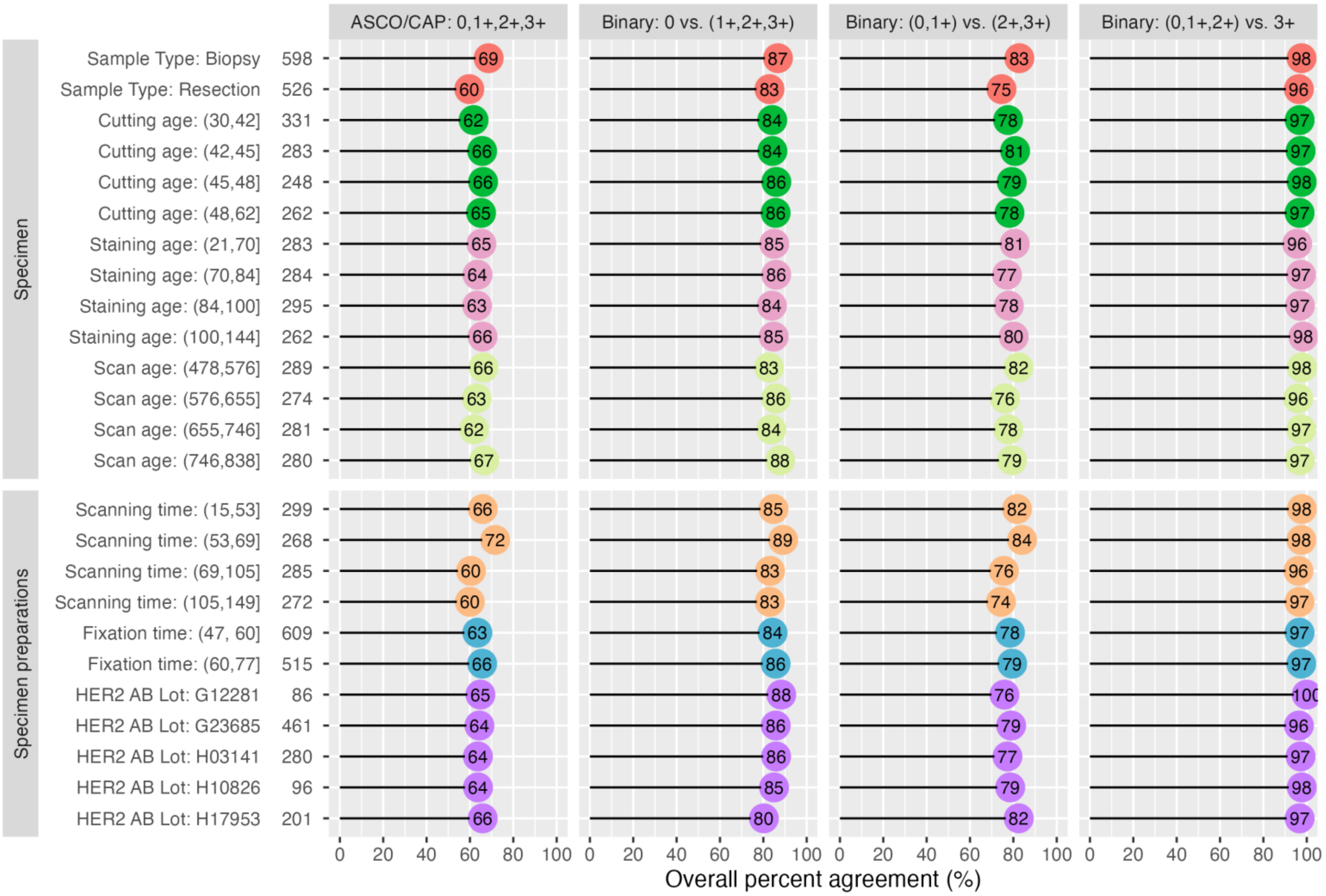
What Drives Variability in HER2 Scoring?



Finding:

Our exploratory analyses suggest that sample type, Ki67, PR, and ER status could be associated with the level of agreement among models.

What Drives Variability in HER2 Scoring?



Finding:
Our exploratory analyses suggest that sample type, Ki67, PR, and ER status could be associated with the level of agreement among models.

Evaluating WSIs to Identify Drivers of Variability

Criterion 1	Criterion 2	Criterion 3
At least **one model scored 0** and **one model scored 3+** on the same sample	**Discordance across pathologists** (at least two HER2 score categories away, e.g., 0 and 2+) AND **Discordance across models** (any discordant calls, does not have to be 2 steps)	All models agree and all pathologists agree, but **models and pathologists do not agree**
32 samples*	17 samples*	4 samples

- A trained pathologist reviewed WSIs and provided a summary of observations
- Tool developers reviewed these images (without knowing their scores/which image they were) and provided updated scoring

*3 samples overlapped from Criteria 1 and 2

Key Observations: Sources of Variability in HER2 Scoring

FRIENDS of CANCER RESEARCH

Artifacts and Sample Quality

- Common issues included staining artifacts, crushed cells, and difficulty visualizing cancer cells
- Benign or DCIS cells exhibited positive staining

Heterogeneous Staining Patterns

- Samples with variable staining intensity across tumor regions
- Particularly impactful for HER2 1+ and 2+ cases

Model and Pathologist Alignment

- Cases with HER2 categories 0 or 3+ showed higher agreement, while HER2 categories 1+ or 2+ had less agreement

Impact of Review Process

- Post-review, agreement among models generally improved when addressing ambiguities, such as artifacts or DCIS staining
- Persistent discordance generally remained in complex cases (e.g., Paget’s disease, cytology samples with sparse tumor cells) highlighting opportunities for further model refinement

Conclusions and Next Steps

Key Findings from the Digital PATH Project

- AI tools demonstrate promise in HER2 scoring with highest agreement for HER2 3+ category
- Variability is more pronounced for HER2 0, 1+ and 2+ categories, which has become increasingly relevant with newer HER2-targeted treatments
- A common dataset enabled robust, rapid comparisons across models, helping identify potential sources of variability and informing best practices

The Role of Reference Data Sets

- Enable transparent evaluation of AI tools
- Provide a foundation for aligning methodologies and identifying variability

Next Steps

- Leverage project findings to propose best practices for AI tool development and validation
- Further explore how reference data sets can be leveraged to support AI tools

FRIENDS of CANCER RESEARCH

Project Partners

4D Path, Inc.

Amgen

AstraZeneca

BostonGene

Bristol Myers Squibb

Caris Life Sciences

Daiichi Sankyo

EMD Serono, Inc.

Emory University

GA Green Consulting LLC

GSK

Indica Labs

Johnson and Johnson Innovative Medicine

Karolinska Institutet

Kulig Consulting

Loxo@Lilly

Lunit

Molecular Characterization Laboratory at Frederick National Laboratory

MD Anderson Cancer Center

Merck and Co., Inc.

National Cancer Institute

Nucleai

Panakeia

PathAI

Patient Advocates

Roche Diagnostics

Sanofi

Tempus AI, Inc.

U.S. Food and Drug Administration

ZAS Hospital

University of North Carolina

Verily

Special Thanks:

Statistical Team

Pedro Torres-Saavedra
Jessica Li
Lisa McShane

Common Data Set Provider

Roberto Salgado
Glenn Broeckx
Frederik Deman

Friends of Cancer Research

Brittany McKelvey
Hillary Andrews
Grace Collins
Jeff Allen