

Supplementary Table S4. DAB intensity-threshold determination: optical-density cutoffs and their reproducibility

One pathologist (S.W.C.), three independent calibrations performed on different tissue-microarray images

(A) Nuclear DAB optical-density cutoffs obtained from the three independent calibrations

Detector	Calibration	t1 (negative 1+)	t2 (1+ 2+)	t3 (2+ 3+)
InstanSeg	a	0.10	0.25	0.40
InstanSeg	b	0.12	0.37	0.57
InstanSeg	c	0.13	0.39	0.61
InstanSeg	Mean (primary, applied)	0.117	0.337	0.527
Legacy	a	0.10	0.23	0.36
Legacy	b	0.10	0.34	0.56
Legacy	c	0.13	0.36	0.58
Legacy	Mean (primary, applied)	0.110	0.310	0.500

(B) H-scores obtained from the three independent calibrations (mean/median across cores; n = 528)

Detector	Calibration a	Calibration b	Calibration c	ICC (95% CI)
InstanSeg	95.1/64.9	74.6/44.1	68.4/38.4	0.954 (0.948–0.960)
Legacy	101.3/71.8	86.3/62.4	74.1/43.6	0.957 (0.950–0.962)

(C) Agreement of H-scores derived from the mean versus the median of the three calibrations

Detector	Mean vs median H-score	ICC	Bias	95% LoA
InstanSeg	78.1/45.5 vs. 74.6/44.1	0.997	+3.51	–5.7 to +12.8
Legacy	86.7/57.2 vs. 86.3/62.4	0.996	+0.42	–13.2 to +14.0

(D) Agreement of the three-tier classification obtained under the three individual threshold sets

Detector	Identical assignment across all three sets	Fleiss' κ	Pairwise quadratic-weighted κ
InstanSeg	451/528 (85.4%)	0.82	0.81–0.96
Legacy	445/528 (84.3%)	0.80	0.78–0.90

(E) Agreement of the three-tier classification obtained with the mean versus the median threshold set

Detector	Identical assignment	κ	Linear-weighted κ	Discordant cores
InstanSeg	519/528 (98.3%)	0.968	0.972	9 (all adjacent-tier)
Legacy	515/528 (97.5%)	0.949	0.954	13 (all adjacent-tier)

CI, confidence interval; ICC, intraclass correlation coefficient; LoA, limits of agreement; κ , Cohen's kappa.

^aCutoffs were derived from reference tumor cells that the pathologist labeled by intensity grade (at least 50 cells per grade), taken as the midpoints of the mean nuclear DAB optical density of adjacent grades. The calibration was repeated three times on different tissue-microarray images, and the component-wise mean of the three cutoff sets was used as the primary threshold set; the three individual sets were retained for sensitivity analysis.

^bThe component-wise mean of the three calibrations was applied as the final threshold set and was used to compute the H-scores in all analyses reported here (InstanSeg, 0.117/0.337/0.527; Legacy, 0.110/0.310/0.500). Rounding to two decimal places (InstanSeg, 0.12/0.34/0.53) was examined as a sensitivity analysis and did not materially change cell-level grading or case H-scores; the rounded values lay within the bootstrap 95% confidence intervals of the data-derived thresholds.

^cThe two detection pipelines use different measurements (Nucleus: DAB OD mean for the legacy detector; DAB: Mean for InstanSeg) and different application functions, so cutoffs are not interchangeable between them.

^dICC (2,1) across the three calibrations quantifies intraobserver reproducibility of threshold setting. The within-core standard deviation of the three H-scores averaged 14.0 (InstanSeg) and 14.1 (Legacy) points on the 0-300 scale. The three threshold sets yielded identical three-tier assignments in 85.4% (InstanSeg) and 84.3% (Legacy) of cores (Fleiss' κ 0.82 and 0.80; pairwise quadratic-weighted κ 0.81–0.96 and 0.78–0.90, respectively).

^eUsing the median rather than the mean of the three cutoff sets produced nearly identical H-scores, indicating that the summary rule had negligible influence on quantification. For each detector, the component-wise median was identical to calibration set b.

^fPanels D and E apply the manual H-score cutoffs (8.0/106.0) as the common categorization rule. Panel D compares the three individual threshold sets with one another; panel E compares the classification obtained with the mean of the three sets against that obtained with the median.

^gThreshold determination was assessed for intraobserver reproducibility only; interobserver variability in threshold setting was not evaluated.