BIOINFO
GIW ISCB-Asia
Seoul · Nov 17–20, 2026

# Predictability Is Not Substitutability

## A cost-of-substitution framework for H&E-based molecular prediction across 5 cancers

**Geongyu Lee**[1,2,*,†], Ka-Kyung Kim[1,*], Sejin Park[1,3], Jaemyun Lyu[1,4], Jeong-Han Seo[1,5,6], Yong Gi Ji[1,7]

[1]Pseudo Lab; [2]OmixAI Co. Ltd.; [3]Seegene Inc.; [4]R&D Center, Genolution Inc.; [5]ROKIT Genomics; [6]Department of Medical Science, University of Ulsan College of Medicine, Asan Medical Center; [7]Quantum C&S
[*]Equal contribution · [†]Corresponding author: rjsrb365@gmail.com

P082
Long abstract
doi.org/10.5281/zenodo.24068874

**Summary.** Discrimination (AUROC) shows that a molecular marker is *predictable* from H&E; it does not show that the prediction can *substitute* for the molecular test. Under one protocol, 10 sealed-forward endpoints in lung, gastric and head-and-neck cancer resolved as **7 candidates: 1 confirmed · 6 undecided** and **3 positive controls**: a near-ceiling AUROC (0.939, lung histology) was excluded by a site audit, and all other actionable endpoints were undecided rather than negative.

## 1. Background

H&E models predict MSI, driver mutations and subtypes at high reported AUROC [1–3], but AUROC does not say which treatment a patient receives when the prediction is wrong [4].

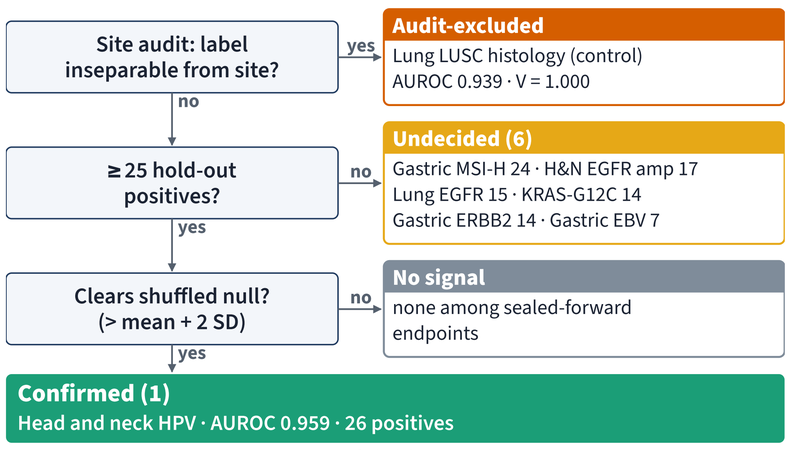
In **our prior work**, an H&E recurrence model reached sensitivity 0.857 for low-risk but 0.529 for high-risk patients [5]. Such predictors are often confounded [6] and drop from AUROC 0.689 to 0.630 externally [7].

## 2. Cost of substitution

We define the cost of substitution as the confusion matrix weighted by treatment distance under a routing fixed in advance:

$$cost = \Sigma_{i,j}\, C_{ij} \cdot D_{ij}$$

The **primary measure** is the distance-free **misassignment rate**. It prices an error at the moment of assignment, not drug response.




Controls: H&N grade passed; gastric Lauren failed; lung histology above. Histology-only baseline was recorded but is not a verdict criterion.

**Verdict rules.** Counts are the 10 sealed-forward endpoints of Figure 2. Gastric ERBB2 is classified Undecided. The histology-only baseline was a reported control, not part of the verdict rules (see Figure 4).

## 3. Methods

- **Cohorts:** breast (TCGA-BRCA anchor, ~1,010 slides), lung, colorectal, gastric, head and neck.
- **Pre-processing:** 256 × 256 tiles at 20×, Otsu segmentation, ≤ 5,000 tiles/slide.
- **Model:** UNI v1 embeddings (1,024-d) [10] + CLAM-SB attention MIL [11].
- **Evaluation:** site-disjoint hold-out only.
- **Controls:** 5-seed label-shuffled null (AUROC > null mean + 2 SD); prevalence baseline; histology-only / pixel-mean baseline (reported, not a verdict criterion); 1,000× bootstrap CI.
- Predictions **sealed by commit** before hold-out results were produced (lung, gastric, head and neck).
- **Scope:** 10 sealed-forward endpoints (incl. 3 positive controls); breast/colorectal (4) and endometrial (1) reported but not counted.

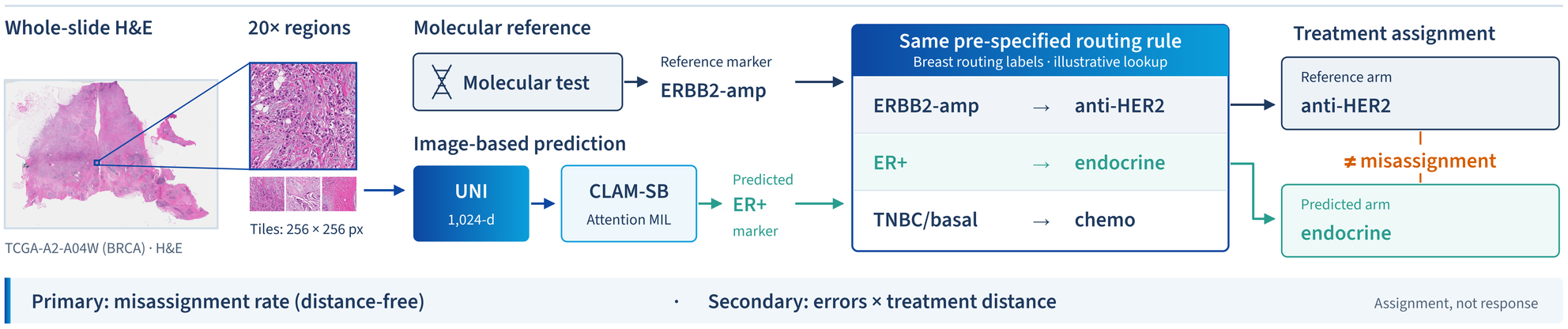



**Figure 1. Framework.** The molecular result and the H&E prediction pass through the same pre-specified treatment routing; a disagreement in the assigned arm is a misassignment. Breast-anchor misassignment: see Figure 5.

## 4. Results

**Table 1.** Sealed-forward hold-out: lung, gastric, head and neck.

| Endpoint | AUROC | 95% CI | Hold-out pos. | Verdict |
|---|---|---|---|---|
| **Head and neck HPV** | 0.959 | 0.921–0.986 | 26 | Confirmed‡ |
| Lung LUSC histology (control) | 0.939 | 0.905–0.967 | 153 | Audit-excluded |
| Gastric MSI-H | 0.860 | 0.759–0.941 | 24 | Undecided |
| Lung EGFR activating | 0.852 | 0.725–0.953 | 15 | Undecided |
| Lung KRAS-G12C | 0.681 | 0.577–0.783 | 14 | Undecided |
| Gastric ERBB2 amplification | 0.644 | 0.523–0.771 | 14 | Undecided |

Also undecided (Figure 2): H&N EGFR amp (17 pos.), gastric EBV (7). Not counted (anchor/retrospective): breast HER2 0.599; colorectal anti-EGFR 0.705 (84 pos.), MSI-high 0.918 (21), BRAF V600E 0.882 (15). ‡ Primary model (UNI). With 20 label shuffles HPV cleared the null under UNI, UNI2-h and Virchow2 [12] (5 shuffles: Virchow2 0.9199 vs 0.9234). Label–site association Cramér's V = 0.397.

**Confirmation.** Head and neck HPV was the only non-control endpoint meeting the primary rule. It has known histological correlates (non-keratinising, basaloid morphology) [8] and HPV association has been predicted from H&E in two cohorts of 273 patients [8].




**Figure 4.** Baseline check: histological subtype alone, with no image, beats the image model for lung KRAS-G12C. Remains Undecided (14 positives): this baseline is a reported control, not a verdict rule, so the contrast is an observation only.

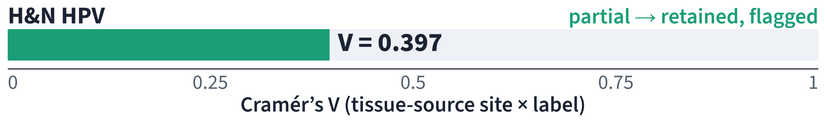



**Figure 3.** Site-confounding audit. In lung, TCGA institution codes coincide exactly with LUAD/LUSC status (V = 1.000) [9], so the 0.939 AUROC was audit-excluded; HPV shows a partial label–site association (V = 0.397).

## 4a. Positive-count ceiling

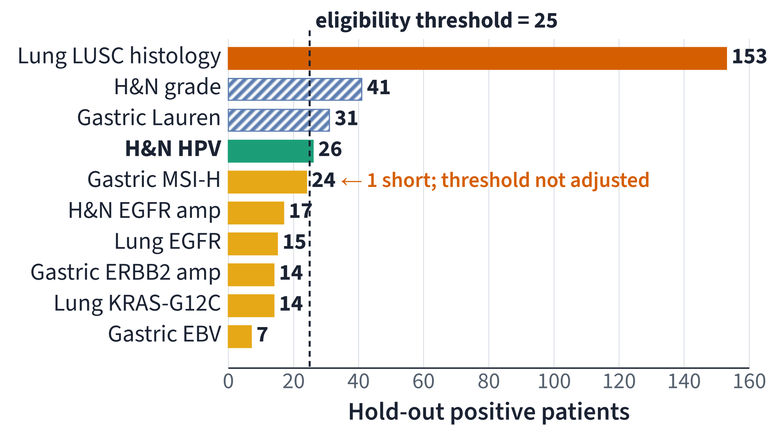



**Figure 2.** Hold-out positives vs the eligibility threshold of 25 (hatched: positive controls). All six non-HPV candidate endpoints fall below it — undecided, not negative. Gastric MSI-H fell one short; the criterion was not adjusted.

## 4b. Breast anchor

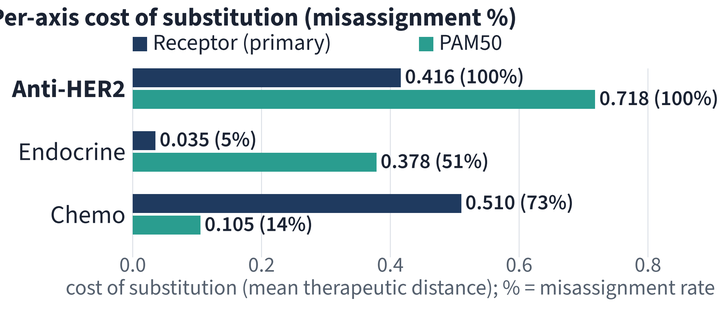



**Figure 5.** Per-axis substitution cost under two pre-specified routings. Every anti-HER2 slide is misassigned under both; endocrine/chemotherapy costs reverse between routings (no safety claim). The anti-HER2 − endocrine contrast excludes 0 under both (receptor 0.381 [0.331–0.427]; PAM50 0.340 [0.251–0.425]). In an external pCR cohort (exploratory), the anti-HER2 prediction did not stratify response (AUROC 0.533).

## 5. Discussion

Retrospective and hypothesis-generating. Site-disjoint splits block leakage, not confounding. Power alone does not explain the pattern (Figure 6).

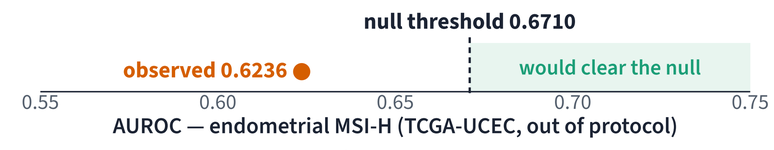



**Figure 6.** Power is not everything. Out-of-protocol endometrial MSI-H met the 25-positive criterion yet failed null exclusion; excluded from the confirmation tally.

## 6. Conclusion

1. Explicit verdict rules separated **1 confirmed** and **6 undecided** candidates, and audit-excluded a near-ceiling control.
2. The framework rejected a near-ceiling AUROC and, under the primary model, retained an externally corroborated endpoint.
3. Reporting the undecidable as **undecidable**, not negative, is the central point; the procedure transfers to any proposed H&E surrogate.

### References

1. Kather JN et al. Nat Med 2019.
2. Coudray N et al. Nat Med 2018.
3. Kather JN et al. Nat Cancer 2020.
4. Vickers AJ, Elkin EB. Med Decis Making 2006.
5. Lee G et al. Sci Rep 2025.
6. Dawood M et al. Nat Biomed Eng 2026.
7. Vanderbilt C et al. arXiv:2603.20848, 2026.
8. Klein S et al. Clin Cancer Res 2021.
9. Howard FM et al. Nat Commun 2021.
10. Chen RJ et al. Nat Med 2024.
11. Lu MY et al. Nat Biomed Eng 2021.
12. Zimmermann E et al. arXiv:2408.00738, 2024.

**Acknowledgement.** This research was supported by the "Open Source AI·SW Developer and Community Support Program" funded by the Ministry of Science and ICT (MSIT), Republic of Korea. Conducted as an independent Pseudo Lab project; corporate affiliations are listed for identification only.

**Code & results**
github.com/Geongyu/predictability-substitutability