PRECISELY WRONG: CONFORMAL PREDICTION FOR HONEST PRECISION ONCOLOGY

C. Stacy[1] and S. Das[1]

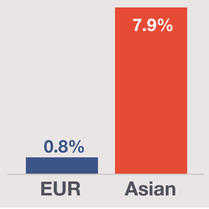
1.School of Pharmacy and Biomolecular Sciences, RCSI University of Medicine and Health Sciences, Dublin, Ireland.

RCSI UNIVERSITY OF MEDICINE AND HEALTH SCIENCES

ASIAN ANCESTRY – STOMACH CANCER

9.9×

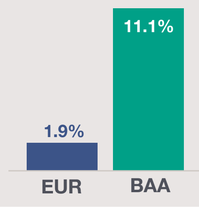
confident calls wrong (vs. European)



AFRICAN ANCESTRY – UTERINE CANCER

5.8×

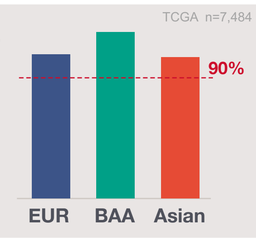
endometrioid misread as aggressive serous



GROUP-CONDITIONAL CONFORMAL

90%

true type in set for every ancestry



Introduction

- DNA methylation signatures can differentiate difficult-to-diagnose cancers.
- Algorithms have been developed to predict cancer classes.
- Selective prediction (abstains when unsure) is widely used.

THE CLASSIFIER (crossNN) *Yuan et al. 2025*

$$P(\mathrm{y} \mid \mathrm{x}) = \mathrm{softmax}(\mathrm{Wx})$$

W: 149×226,085 CpGs
Single-layer neural network, 99.75% masking of binarized probes

- Assumes confidence corresponds to correctness of prediction.
- But, classifiers usually trained on ancestrally-skewed datasets
  - (TCGA ~80% European ancestry; Fig. 1) *Carrot-Zhang et al, 2020.*
- Is the **reliability** of confident calls **ancestry-dependent?**

Objective

- Audit methylation classifier for ancestry-dependent reliability of confident classification calls
- Test whether group-conditional conformal restores per-ancestry calibration.

Methods & Data

- TCGA HM450 methylation. *N*=7,484 cases (EUR, BAA, Asian).
- 5-fold out-of-fold logits. APS $\alpha = 0.10$ *(Romano et al. 2020).*

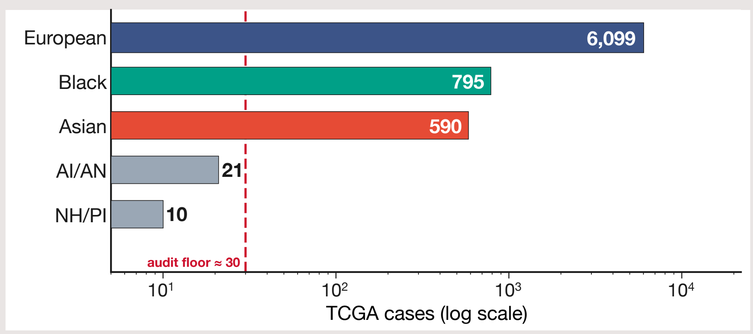


Figure 1. The methylation reference cohort (TCGA) is overwhelmingly European ancestry. Per self-reported ancestry group (log scale): European 6,099 (~80%), African 795, Asian 590. American Indian/Alaska Native AI/AN (21), and Native Hawaiian/Pacific Islander NH/PI (10) fall below the ~30-case audit floor needed to audit any single cancer (dashed line) and are excluded from subgroup testing.

Unequal reliability of confident calls

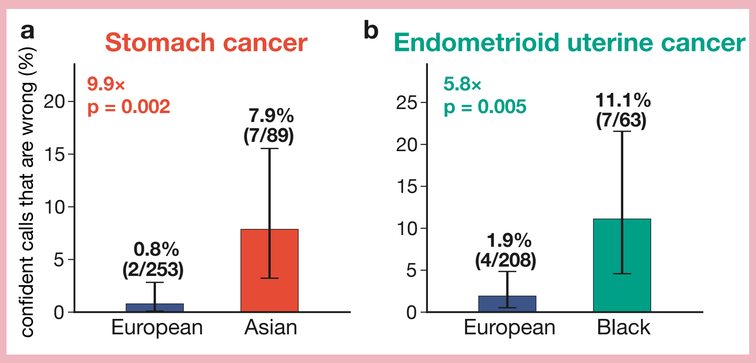


Figure 2. Within a single cancer type, confident calls fail far more often for under-represented ancestries. Observed error rate among high-confidence predictions (calibrated probability ≥0.8); bars are 95% Cis. **(a)** Stomach cancer: Asian ancestry 7.9% vs European ancestry 0.8% (9.9×). **(b)** Endometrioid uterine cancer: African ancestry 11.1% vs European 1.9% (5.8×). Across a 28 cell ancestry-cancer type screen, gastric cancer gap survives Bonferroni correction; the endometrioid gap survives FDR at α=0.05.

- These classification errors are wrong but confident.
- Disproportionately impacts cancers with known existing disparities: endometrial cancer has among largest racial mortality gaps in oncology.
- Disparities persist in single cancer types; case-mix (which cancers each group gets) cannot explain observed differences (Fig. 2).

Limits of threshold-based abstention

- Abstention can only act on model confidence. If the model is confidently incorrect, then results are reported as correct.
- High-confidence errors are not identified by confidence-thresholds alone (Fig. 2).
- Raising confidence thresholds increases abstention rates on less represented populations without resolving the errors.

Molecular basis of misclassification

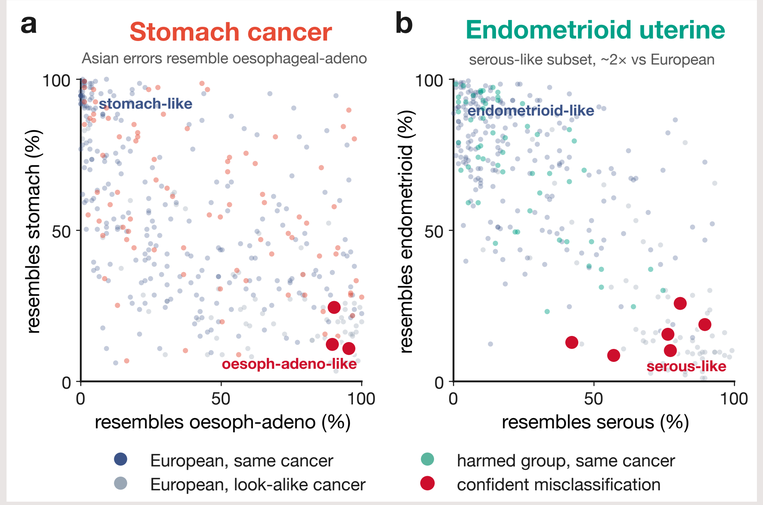


Figure 3. Misclassified tumours genuinely carry the methylation signature of the "look-alike" cancer. Each point represents one tumour, scored by how closely its methylation pattern matches markers of its own cancer type (y-axis) versus the "look-alike" type (x-axis). Confident misclassifications are enlarged with red points. **(a) S**tomach cancers in individuals with Asian ancestry misread as oesophageal adenocarcinoma. **(b)** Endometrioid uterine cancers in individuals with African ancestry misread as the aggressive serous subtype.

- Confidently misclassified tumours genuinely match the look-alike subtype (Fig. 3)
- High-confidence misclassifications are ~2× more likely in people of non-European ancestry. Matches known excess of aggressive endometrial disease in women of African ancestry
- crossNN model's own weight attribution elucidates biology
- ***GRIA4*** identified as top driver by model, recently validated as a methylation marker for endometrial cancer. *(Chen et al., 2026)*

Confounding in the fairness audit

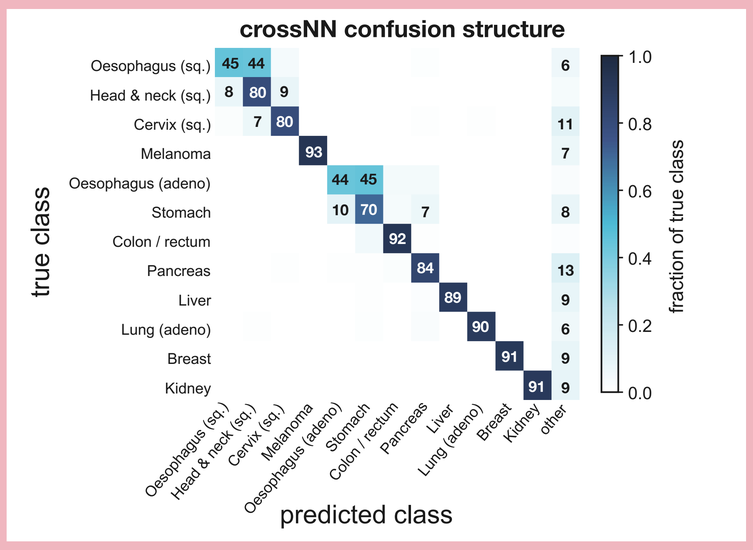


Figure 4. Misclassifications between cancer types are rarely equally bidirectional. Row-normalized crossNN confusion matrix (each row is true class, colour is the % predicted as each class). Squamous-cell (sq) and adenocarcinoma (adeno) oesophageal cancer are classified separately. Twelve classes shown; other class predictions included in right-most column.

- Common classification errors are directional: a rarer cancer is absorbed into a similar class. Squamous-oesophagus reads as head-and-neck 44% of the time and adeno-oesophagus as stomach 49%, while the reverse is <10%.
- Confusions are between cancer types, so accuracy-by-ancestry requires correcting for prevalence of classes between ancestral groups in the training dataset.

Group-conditional conformal prediction

NONCONFORMITY SCORE (APS) *Romano et al. 2020*

$$\mathrm{s}(\mathrm{x}, \mathrm{y}) = \sum_{k:\, \pi_k \geq \pi_y} \pi_k$$

CALIBRATE WITHIN EACH ANCESTRY

$$\hat{q}_a = \mathrm{Quantile}_{\lceil (\mathrm{n}_a+1)(1-\alpha) \rceil / \mathrm{n}_a}\{S_i : A_i = a\}$$

PREDICTION SET FOR A NEW PATIENT

$$C(\mathrm{X}) = \{y : s(x, y) \leq \hat{q}_a\}$$

PER-GROUP COVERAGE GUARANTEE

$$P(Y \in C(\mathrm{X}) \mid \mathrm{A} = a) \geq 1 - \alpha$$

x : methylation profile
y : candidate tumour type
A, a : Ancestry / group
W : crossNN weights
π : softmax(Wx)
n_a : calibration cases
k : class k
$\mathrm{s}(x, y)$: nonconformity score

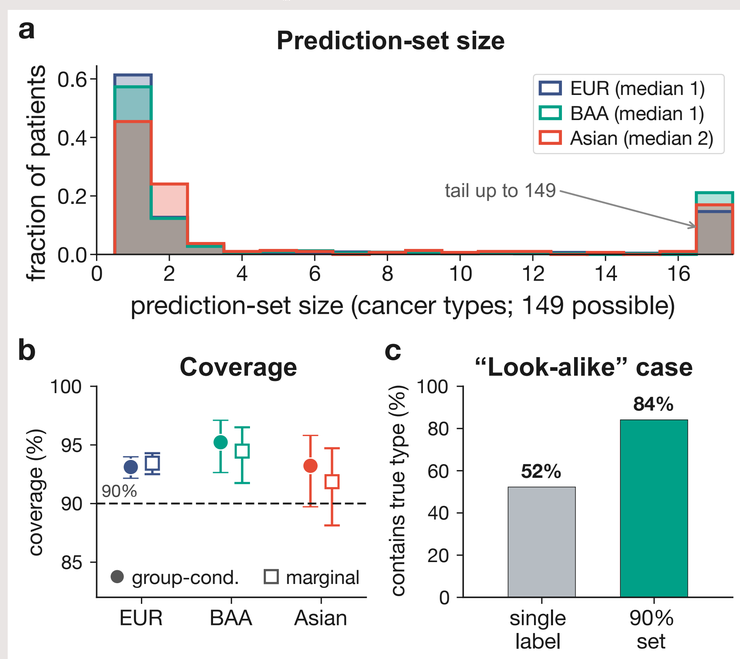


Figure 5. Group-conditional conformal returns calibrated sets that stay small for easy-to-classify cases and widen for hard cases. (a) Set size: most patients are assigned 1 or 2 classes of 149 classes across all ancestries examined with a long right tail. **(b)** Coverage is ~90% within every ancestry group (dashed target line). 95% Cis are wide where sample group sizes are small. Shapes show estimated group-conditional (circle) and marginal (square) true-class coverage **(c)** For the squamous-oesophageal look-alike case, the 90% set contains the true type 85% of the time, versus 52% for the single top-1 label.

- Conformal prediction outputs calibrated set that holds true type ~90% of cases (Fig. 5).
- Confident cases stay one class; ambiguous tumours provide appropriately cautious findings.
- Calibration within ancestry results in ~90% coverage per ancestry group, even with cohort shift (*Lu et al. 2021*)

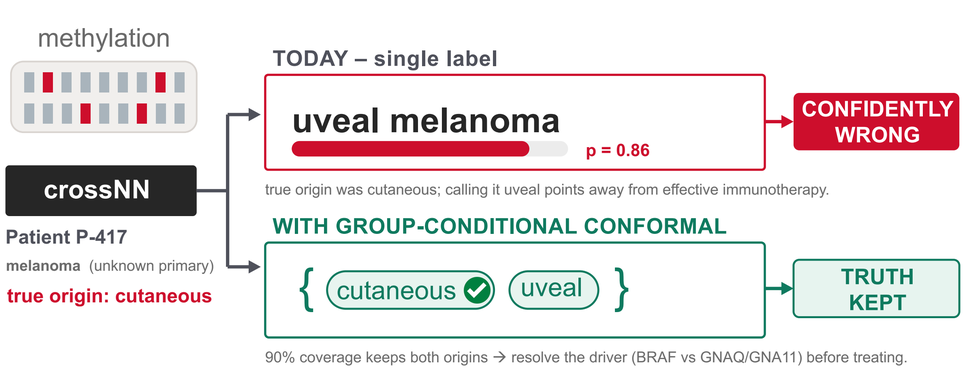
Conclusions

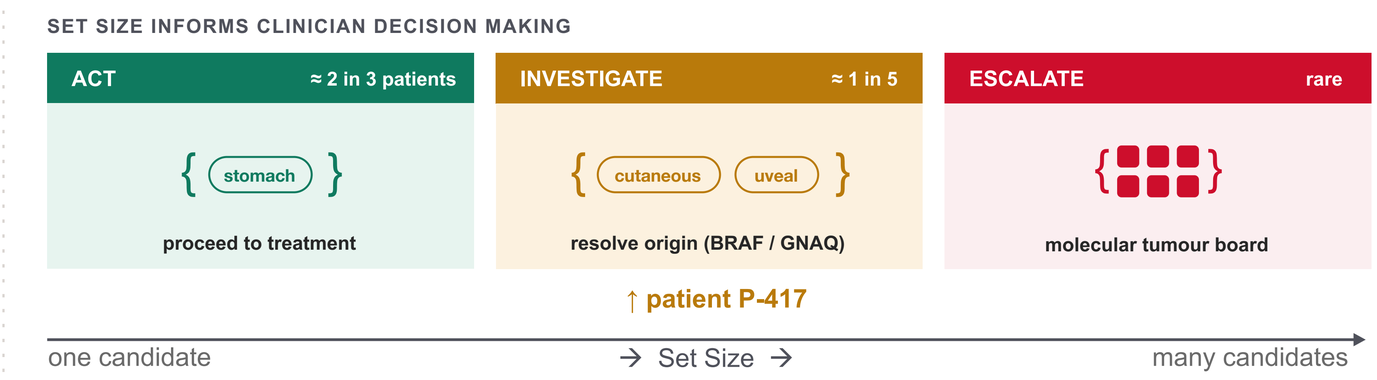
- Differences in molecular classifier performances are observed when comparing across ancestral groups.
- Group-conditional conformal makes the model's uncertainty more honest per ancestry, supporting replacing of abstention with widened sets.
- *Limitations*: coverage determined out-of-fold; ancestry is self-reported race.

References

1. Yuan D, *et al*. crossNN is an explainable framework for cross-platform DNA methylation-based classification of tumors. *Nature Cancer*. 2025;6(7):1283–1294.
2. Romano Y, Sesia M, Candès EJ. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems (NeurIPS)*. 2020;33:3581–3591.
3. Lu C, Lemay A, Chang K, Höbel K, Kalpathy-Cramer J. Fair conformal predictors for applications in medical imaging. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022;36(11):12008–12016.
4. Carrot-Zhang J, *et al*. Comprehensive analysis of genetic ancestry and its molecular correlates in cancer. *Cancer Cell*. 2020;37(5):639–654.
5. Chen *et al*. Development and validation of hypermethylated gene markers in cervical cytological samples for detecting endometrial cancer (EndoMethy-I). *Frontiers in Oncology*. 2026.

CLINICAL TRANSLATION





Acknowledgements Health Research Board TRANSCAN

Supported by the Health Research Board under SPELCASTER (TRANSCAN-3 ERA-NET Cofund, EU). Das Lab, RCSI.

WWW.RCSI.COM