

Additional file 5. Robustness analyses and source-data guide

Patient-level multimodal LSIL-or-above risk stratification

15 July 2026

1 Scope

This appendix summarizes prespecified or fixed-protocol sensitivity analyses that were not used to select a new headline model. The accompanying workbook contains aggregate machine-readable results only. It excludes patient identifiers, raw images, filenames, patient-level predictions and embedding arrays.

2 Fully out-of-fold image representations

The fully out-of-fold analysis replaced the cached image-representation PCA branch. Each training patient’s representation came from a ConvNeXt checkpoint trained without that patient’s group fold; validation and test representations used only checkpoints that had not trained on those patients. Scaling and 32-component PCA were fitted on training patients only.

Analysis	ROC-AUC	PR-AUC	Brier	BA	Interpretation
Diagnostic-context model	0.863	0.762	0.185	0.838	Original internally evaluated estimate
Fully OOF representation	0.814	0.619	0.203	0.712	Conservative representation sensitivity

The paired ROC-AUC difference was -0.048 (95% CI -0.138 to 0.038; P=0.274), and the PR-AUC difference was -0.143 (-0.249 to 0.030; P=0.127). The Brier-score difference was 0.018 (0.008 to 0.028; P=0.001). These data support prominent reporting of attenuation rather than a claim that the cached representation was fully leakage-free.

3 Annotation-training ablation

Raw-only training used three raw-image epochs per fold. Annotation-informed training used one annotation-overlay pretraining epoch followed by two raw-image fine-tuning epochs. Validation and test inference always used raw images.

Protocol	ROC-AUC	PR-AUC	Brier	BA	Interpretation
Raw-only	0.793	0.692	0.285	0.763	Fixed sensitivity comparator
Annotation pretraining + raw fine-tuning	0.779	0.644	0.251	0.695	No discrimination advantage established

The annotation-minus-raw ROC-AUC difference was -0.014 (95% CI -0.148 to 0.109; P=0.846), and the PR-AUC difference was -0.048 (-0.180 to 0.102; P=0.594).

4 Duplicate-conservative eligibility

Duplicate components were defined by exact SHA-256 identity, exact perceptual-hash identity, or cross-patient perceptual-hash Hamming distance of four or less. The primary and duplicate-conservative rows below use different eligible subsets and are not paired same-sample comparisons.

Analysis	N	Events	ROC-AUC	PR-AUC	BA	F1	Role
Primary all-available-image	51	16	0.863	0.762	0.838	0.757	Primary held-out test
Duplicate-conservative	38	11	0.758	0.731	0.753	0.640	Eligibility sensitivity

5 Group OOF, source-batch and deployable-variable analyses

These analyses used a separate fixed supplementary CatBoost configuration: 300 iterations, depth 5, learning rate 0.025, L2 leaf regularization 10, random strength 1, bagging temperature 0 and balanced automatic class weights. They were not used to replace the 77-feature diagnostic-context model.

Analysis	Feature boundary	N	Events	ROC-AUC	PR-AUC
Five-fold stratified group OOF	Full 62-feature diagnostic context	356	109	0.872	0.688
Five-fold stratified group OOF	Deployable clinical + colposcopy scores	356	109	0.692	0.491
Five-fold stratified group OOF	Deployable 56-feature multimodal	356	109	0.686	0.441
Held-out source Batch 1	Full 62-feature diagnostic context	184	57	0.791	0.572
Held-out source Batch 2	Full 62-feature diagnostic context	172	52	0.722	0.473

Source-batch holdout is internal transportability analysis, not external validation. The deployable-variable attenuation shows that contemporaneous current-test and diagnosis-proximal variables account for substantial discrimination.

6 ZhengHou paired analyses under deployable boundaries

Adding ZhengHou features to deployable clinical variables changed ROC-AUC by -0.021 (95% CI -0.062 to 0.021; $P=0.324$). Adding ZhengHou/BGE semantics to deployable clinical plus colposcopy features changed ROC-AUC by -0.006 (-0.047 to 0.035; $P=0.757$). These paired intervals do not establish an independent semantic increment.

7 Complete-case selection

The 356 included complete-multimodal patients were younger than the 499 excluded labelled records (median 38 versus 41 years; standardized mean difference -0.238). Source-label and examination-period distributions differed (Cramer's V 0.227 and 0.238). The positive-outcome proportions were 30.6% and 36.5%, respectively (Cramer's V 0.061; $P=0.075$). Complete-modality eligibility may therefore have introduced selection bias.

8 Interpretation

The primary ROC-AUC of 0.863 is best treated as an internal diagnostic-context estimate. Fully OOF representations, duplicate-conservative eligibility, source-batch holdout and removal of diagnosis-proximal variables all expose lower-performance conditions. The results support transparent internal validation and a frozen prospective external study; they do not support deployment, state-of-the-art superiority or a statistically conclusive ZhengHou improvement.