Original Paper

Clinical Plausibility and Pretraining Utility of Virtual Narrow-Band Imaging for Laryngeal Lesion Classification: Multicenter Model Development and Validation Study

Jie-Lin Huang[1], MD; Li-Juan Li[2], MD; Mei-Ling Wang[3], MD; Ji-Qing Zhu[1], MD; Li-Zhou Dou[1], MD; Yu-Meng Liu[1], MD; Yu-Da Zhao[1], MD; Xue Zhang[1], MD; Xiao-Guang Ni[1], MD, PhD

[1]Department of Endoscopy, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[2]Department of Otorhinolaryngology, The People's Hospital of Wenshan Prefecture, Wenshan, Yunnan, China

[3]Department of Endoscopy, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, China

Corresponding Author:

Xiao-Guang Ni, MD, PhD
Department of Endoscopy, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital
Chinese Academy of Medical Sciences and Peking Union Medical College
No. 17 Panjiayuan South Lane, Chaoyang District
Beijing 100021
China
Phone: 86 10 8778 7606
Email: nixiaoguang@126.com

Abstract

Background: Pathology-labeled optical narrow-band imaging (NBI) is difficult to assemble for laryngeal lesion classification. Translating existing white-light imaging (WLI) into virtual NBI (Vir-NBI) may improve learning from limited real-NBI data, but its value beyond that of the original images is uncertain.

Objective: This study aimed to determine whether Vir-NBI pretraining improves real-NBI classification or label efficiency compared with pretraining on the same pathology-labeled WLI.

Methods: This multicenter study compared ResNet-50 classifiers pretrained on pathology-labeled WLI or paired Vir-NBI images generated by a frozen, region-of-interest–guided CycleGAN (cycle-consistent generative adversarial network). The primary outcome was the image-level area under the receiver operating characteristic curve (AUC) difference (Vir-NBI minus WLI). Label efficiency was evaluated in these external cohorts with 10 to 100 labeled real-NBI patients and 25 paired training runs per sample size. Color-matched augmentation and paired test-modality analyses examined the influence of color change and test modality. Prospective temporal validation and a 12-clinician assessment evaluated patient-level performance and perceived image quality.

Results: The CHCAMS (National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College) cohort comprised 1245 patients with 8476 paired WLI–Vir-NBI images. External validation included 182 patients from CHSZCAMS (Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College) and 129 from People's Hospital of Wenshan Prefecture (WSPH). External image-level AUCs were 0.806 with WLI and 0.787 with Vir-NBI (difference –0.019, 95% CI –0.047 to 0.008; *P*=.17). With 10 labeled real-NBI patients, external patient-level AUCs were 0.843 and 0.773, respectively (difference –0.070, 95% CI –0.116 to –0.029; Holm-adjusted *P*=.006). The normalized area-under-the-learning-curve difference across 10 to 100 patients favored WLI (–0.049, 95% CI –0.090 to –0.011; Holm-adjusted *P*=.01). Color-matched WLI augmentation yielded higher external image-level AUC than with Vir-NBI (0.774 vs 0.739; difference 0.036, unadjusted 95% CI 0.004-0.070). In the 58-case examination-linked cohort, relative Vir-NBI performance improved on translated test images (primary-model interaction 0.235, unadjusted 95% CI 0.080-0.390); fixed-duration and all-view sensitivity analyses had case-level CIs that included 0. In the CHCAMS prospective validation cohort (138 patients), patient-level AUCs were 0.846 (95% CI 0.775-0.910) and 0.844 (95% CI 0.777-0.905), respectively. WLI identified 90.9% (60/66) of high-risk patients, compared with 81.8% (54/66) with Vir-NBI, reflecting 6 discordant patients (exact McNemar *P*=.03; Holm-adjusted *P*=.06). Among 100 image pairs from the CHCAMS source cohort,

clinicians rated plausibility as 4 to 5 in 86.9% (1043/1200) of ratings; intraclass correlation coefficients for the mean ratings of 12 clinicians ranged from 0.398 (95% CI 0.265-0.497) to 0.583 (95% CI 0.471-0.659).

Conclusions: Vir-NBI pretraining did not demonstrate incremental value over WLI for real-NBI classification. External learning curves favored WLI with small real-NBI training sets. Test-domain dependence suggests that compatibility with real-NBI may contribute to performance differences.

JMIR Med Inform 2026;14:e109811; doi: 10.2196/109811

Keywords: narrow-band imaging; laryngoscopy; synthetic data; image-to-image translation; label efficiency; external validation; deep learning; medical imaging; CycleGAN

Introduction

Developing clinically reliable classifiers for optical narrow-band imaging (NBI) requires sufficiently large and diverse collections of pathology-labeled real-NBI images, but such data are difficult to assemble across institutions. Optical NBI enhances superficial mucosal and vascular patterns that are less apparent under conventional white-light imaging (WLI), and meta-analyses support its diagnostic value in detecting malignant transformation and laryngeal leukoplakia [1-3]. However, access to NBI systems remains uneven, and only a subset of NBI examinations can be linked to definitive histopathology. This scarcity of pathology-labeled real-NBI data creates a target-modality bottleneck for classifier development and external validation [4,5]. Studies of laryngeal leukoplakia have also emphasized classification criteria, reader training, and the need for histopathologic confirmation [6-8]. Recent studies have applied deep learning to laryngoscopic and NBI image classification [9-11]. Related work has also explored real-time detection and boundary delineation [12,13].

Virtual NBI (Vir-NBI) offers an appealing potential solution. By translating existing pathology-labeled WLI into an NBI-like appearance, Vir-NBI could enable established WLI archives and their pathology labels to be reused for target-modality model development, potentially reducing the amount of labeled real-NBI data required. Our group previously developed a region-of-interest–guided CycleGAN (cycle-consistent generative adversarial network) that deterministically translates laryngeal WLI into Vir-NBI and evaluated its technical image quality and reader-facing performance [14]. Cycle-consistent adversarial translation is particularly attractive when exactly registered images across modalities are unavailable [15]. However, translation does not create new patients, pathology labels, or target-modality acquisitions; it re-encodes information already present in the source WLI. Any benefit must therefore arise from producing representations that are more useful for learning from real-NBI. The broader medical-imaging literature frames these approaches as synthesis, modality translation, or domain adaptation [16-18].

Synthetic medical imaging may help address data scarcity [16,19], but fidelity failures, label-associated artifacts, and distributional biases can impair downstream performance [20,21]. Similarly, reducing an apparent domain gap does not necessarily preserve pathology-discriminative information [17,18]. A translated image may resemble real-NBI while suppressing lesion features or introducing shortcuts that are unhelpful for classification. Evaluations based only on visual quality, distributional similarity, or standalone model performance therefore cannot determine whether translation adds value beyond the original source images. The appropriate counterfactual is direct pretraining on the same pathology-labeled WLI under otherwise identical conditions. The need to test generalization beyond the development setting has been emphasized in imaging AI evaluation [22-24], and broader clinical AI reviews likewise identify external validation and transportability as central requirements [25]. If Vir-NBI contributes useful target-modality information, it should improve external discrimination or reduce the amount of labeled real-NBI required relative to matched WLI pretraining.

We therefore conducted a multicohort model development and validation study comparing 2 matched strategies: direct pretraining on pathology-labeled WLI and pretraining on corresponding Vir-NBI translations generated from the same source patients and labels. Both strategies subsequently underwent real-NBI fine-tuning with shared optimizer settings. We hypothesized that Vir-NBI pretraining would improve real-NBI label efficiency and external discrimination relative to matched WLI pretraining. We further separated the effects of virtual translation and real-NBI fine-tuning, examined learned representations and clinician-rated image quality, and evaluated model transport in the independent, prospectively collected, biopsy-indicated CHCAMS (National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College) validation cohort. Together, these analyses tested not merely whether Vir-NBI appeared clinically plausible but also whether it addressed the real-NBI data scarcity that motivated its use.

Methods

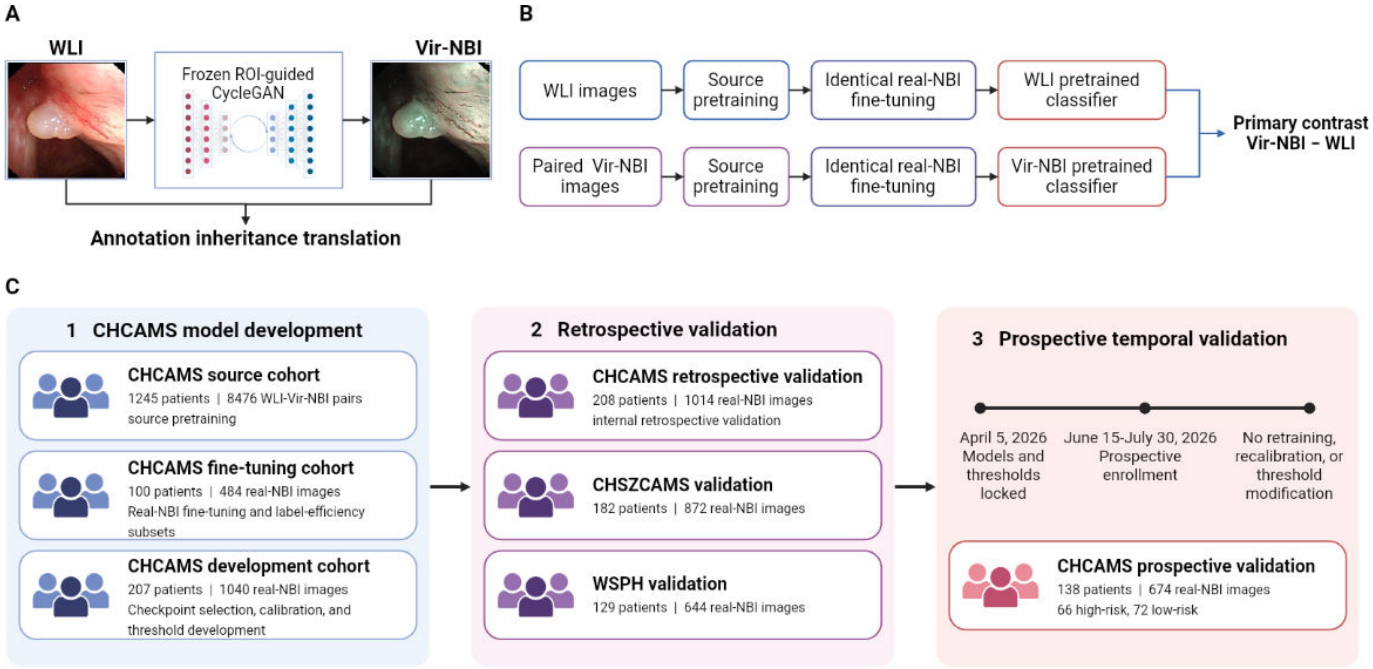
Study Design and Reporting

This multicenter model development and validation study evaluated whether Vir-NBI provided incremental training value over pathology-labeled WLI for real-NBI classification of laryngeal lesions. The study comprised retrospective model development and internal validation at the CHCAMS, retrospective external validation at CHSZCAMS (Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College), and the People's Hospital of Wenshan Prefecture (WSPH) under

established external-validation principles [26], and prospective temporal validation at CHCAMS.

The primary end point was the image-level area under the receiver operating characteristic curve (AUC) difference averaged across the CHSZCAMS and WSPH external validation cohorts. Cohort assignments were made at the patient level, and resampling accounted for images clustered within patients. The statistical analysis framework is detailed in Section S10 in Multimedia Appendix 1, the statistical analysis plan is provided in Section S14 in Multimedia Appendix 1. The study is reported with reference to TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI; Checklist 1), CLAIM (Checklist for Artificial Intelligence in Medical Imaging; Checklist 2) 2024, and STARD-AI (Standards for Reporting Diagnostic Accuracy–AI; Checklist 3) [27-29]. Figure 1 presents the study design.

Figure 1. Study design and validation cohorts. (A) Frozen region of interest (ROI)–guided CycleGAN (cycle-consistent generative adversarial network) translation of white-light imaging (WLI) into paired virtual narrow-band imaging (Vir-NBI), retaining patient-level pathology labels. (B) WLI and Vir-NBI pretraining with shared real narrow-band imaging (real-NBI) fine-tuning cohorts, architecture, and evaluation; the primary contrast was Vir-NBI minus WLI. (C) Cohort roles, patient counts, and temporal-validation timeline. CHCAMS: National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College; CHSZCAMS: Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College; WSPH: People's Hospital of Wenshan Prefecture.



Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of CHCAMS (approval number 22/454-3656). The institutional review board waived the requirement for informed consent for the retrospective cohorts. The prospective validation component was registered with the Chinese Clinical Trial Registry before enrollment (ChiCTR2600126503) and was conducted as a noninterventional observational study. Model predictions were not disclosed to patients or clinicians and did not influence endoscopic examination, biopsy, pathologic assessment, treatment, or any other aspect of clinical care.

Study Population and Cohorts

Patients were eligible if they underwent laryngoscopic examination at a participating institution, had at least one evaluable WLI or NBI image of a laryngeal lesion, and had a definitive histopathologic diagnosis. No patient contributed to more than one model-development or validation cohort.

Four retrospective CHCAMS cohorts were used for source-image pretraining, real-NBI fine-tuning, model development, and internal validation. The CHCAMS source cohort comprised 1245 patients with 8476 pathology-labeled WLI images; each WLI image had 1 corresponding Vir-NBI translation. The CHCAMS real-NBI fine-tuning cohort of 100 patients with 484 real-NBI images was used for target-modality fine-tuning and label-efficiency experiments. Model selection, probability calibration, and operating-threshold selection were performed in the CHCAMS development cohort of 207 patients with 1040 real-NBI images. The CHCAMS retrospective internal validation cohort comprised 208 patients with 1014 real-NBI images.

Retrospective external validation was performed in the CHSZCAMS external validation cohort, comprising 182 patients with 872 real-NBI images, and the WSPH external validation cohort, comprising 129 patients with 644 real-NBI images. Neither external cohort was used for model selection, calibration, threshold selection, or hyperparameter optimization. The CHCAMS prospective validation cohort was recruited from June 15 to July 30, 2026.

Reference Standard and Outcome

The reference standard was the final histopathologic diagnosis of the corresponding laryngeal lesion. The primary outcome was low-risk vs high-risk pathology according to the fifth edition of the World Health Organization classification [30,31]. Inflammatory lesions, hyperplasia, polyps, edema, cysts, granuloma, papilloma, tuberculosis, and mild dysplasia were classified as low risk. Moderate dysplasia, severe dysplasia, and invasive carcinoma were classified as high risk.

Vir-NBI Generation

Vir-NBI was generated using a frozen region of interest (ROI)–guided CycleGAN with a 9-residual-block ResNet generator, instance normalization, and PatchGAN discriminators [14,15]. The WLI cycle-reconstruction loss assigned weights of 11 to lesion pixels and 1 to background pixels. Translation used red, green, and blue images resized and padded to 256×256 pixels; ROI masks were not required. Architecture and preprocessing details are provided in Multimedia Appendix 1.

Classifier Development and Training Strategies

All classifiers used ImageNet-initialized ResNet-50 with 224×224 red, green, and blue inputs and a binary output [32,33]. The backbone and classification head were fully trainable. Adaptive moment estimation with decoupled weight decay source pretraining used a learning rate of 0.0001; real-NBI fine-tuning used a learning rate of 0.00001. Both strategies used a maximum of 100 epochs, with early stopping after 5 consecutive epochs without improvement in validation performance. Training used binary cross-entropy with logits, with class weighting during fine-tuning. Models were selected by AUC in held-out CHCAMS source-cohort patients after pretraining and in the CHCAMS development cohort after fine-tuning. Independent validation cohorts were not used for model selection. WLI and Vir-NBI pretraining used images from the same CHCAMS source-cohort patients with the same pathology labels. Both strategies shared subsets from the CHCAMS real-NBI fine-tuning cohort, architecture, learning rate, optimizer, and augmentation. Further training details are provided in Multimedia Appendix 1.

Predictions were averaged across 5 independently trained models. Comparators included direct real-NBI training from an ImageNet initialization, simple framework for contrastive learning of visual representations pretraining followed by supervised real-NBI fine-tuning [34], joint WLI-Vir-NBI pretraining, and strongly color-augmented WLI pretraining. A 2×2 analysis compared WLI and Vir-NBI pretraining with and without real-NBI fine-tuning.

Gradient-weighted class activation mapping attribution was summarized by the proportion of nonnegative intensity within the lesion ROI and the distance between the intensity-weighted saliency centroid and ROI centroid, normalized by the image diagonal. Source-image ROI polygons were transformed for WLI and Vir-NBI. These descriptive measures were potentially dependent on the generator-guidance annotations (Multimedia Appendix 1).

Color and Test-Domain Controls

To assess the contribution of color-change magnitude, WLI color augmentation was calibrated to match the mean color difference in CIELAB space (ΔE76) between paired WLI and Vir-NBI training images from the CHCAMS source cohort. ColorJitter parameters were estimated and verified using separate source-training samples of 1024 images each. Standard WLI, color-augmented WLI, and Vir-NBI models used the same 5 random seeds and were evaluated after 10 source-pretraining and 10 real-NBI fine-tuning epochs. Matching was limited to mean color difference and did not include the full distribution or the direction of color changes (Section S11 and Tables S13-S14 in Multimedia Appendix 1).

To assess whether relative classifier performance depended on the test modality, an examination-linked test-modality cohort was assembled with WLI and real-NBI images from 58 cases (25 low-risk and 33 high-risk). WLI images were translated using the frozen generator. The close-view comparison included 587 translated images and 458 real-NBI images from the same cases. Views were selected from image metadata before the AUC calculation; corresponding images were not registered frame pairs. Both models were trained for 10 source-pretraining and 10 fine-tuning epochs, and the models used in the primary analysis were evaluated on these image sets without refitting. Case probabilities were averaged across images and the 5 models. The test-domain interaction was defined as the Vir-NBI-minus-WLI pretraining AUC difference on translated images minus that on real-NBI. An additional analysis included all 910 translated WLI images. Case selection and overlap exclusions are detailed in Section S12 and Tables S15-S16 in Multimedia Appendix 1.

Label-Efficiency Analysis

Learning curves were constructed using nested, class-balanced training sets of 10, 20, 40, 60, 80, and 100 labeled patients from the CHCAMS real-NBI fine-tuning cohort. The combination of 5 patient-subset sequences and 5 model seeds resulted in 25 paired runs for each training-set size. Models selected using the CHCAMS development-cohort AUC were evaluated in the CHSZCAMS and WSPH external validation cohorts without further training or model selection. Patient-level predictions were obtained by averaging image probabilities, and performance at each training-set size was summarized as the mean patient-level AUC across the 25 runs.

For learning curves in the CHSZCAMS and WSPH external validation cohorts, 10,000 bootstrap replicates resampled patients within each center and risk stratum, subset sequences, and model seeds. Sequence and seed draws were shared across arms, budgets, and centers; external macro-averages were assigned equal center weights. This resampling accounted for dependence among the 25 crossed runs. The normalized area under the learning curves (AULCs) used trapezoidal integration over 10 to 40 and 10 to 100

patients. Holm correction was applied across the 6 training-set-size comparisons and separately across the 2 AULC intervals within each cohort or macroaverage. Center-specific comparisons were not adjusted jointly across centers. CHCAMS development-cohort analyses used the same 3-axis bootstrap approach.

Evaluation End Points

The primary end point was the uncalibrated image-level AUC difference between the Vir-NBI–pretrained and WLI-pretrained strategies, defined as Vir-NBI minus WLI and calculated as an equal-weighted macroaverage across the CHSZCAMS and WSPH retrospective external validation cohorts. CIs and *P* values were estimated using paired, patient-clustered bootstrap resampling. A patient-level external macroaverage, calculated after averaging image probabilities within each patient, was a key secondary analysis.

For the CHCAMS prospective validation cohort, strategy-specific logistic calibration and thresholds were estimated in the CHCAMS development cohort, while maximizing specificity, subject to a development sensitivity of at least 0.90. Calibration functions and thresholds remained fixed for prospective evaluation. Sensitivity and specificity differences were evaluated using paired bootstrap inference and exact McNemar comparisons of paired correctness. Exact-test *P* values were Holm-adjusted across sensitivity and specificity.

Clinician Quality Assessment

A total of 12 clinicians independently assessed 100 WLI–Vir-NBI image pairs from 100 CHCAMS source-cohort patients (50 low-risk and 50 high-risk). WLI served as an unscored morphologic reference, and Vir-NBI was the scored target. Pathologic diagnoses, model predictions, and other clinicians' ratings were concealed. Clinical plausibility, diagnostic confidence, freedom from boundary distortion/oversmoothing, and freedom from vascular artifacts were rated on a 1 to 5 scale, with higher scores indicating better performance. For the 2 artifact-related items, a score of 5 indicated almost no corresponding defect. Boundary distortion and oversmoothing were assessed as a single item.

Score distributions and proportions of ratings of 4 to 5 were summarized descriptively. CIs were estimated using 10,000 crossed-bootstrap resamples of patients and clinicians. Two-way, absolute-agreement intraclass correlation coefficients (ICCs) were calculated for a single clinician and the mean of 12 clinicians, using 100-patient panels for each item and patient-bootstrap CIs with the 12 clinicians held fixed. Associations of patient-level mean ratings with appearance and downstream-utility gains were evaluated using Spearman correlations, crossed-bootstrap CIs, 10,000 patient-label permutations for 2-sided *P* values, and Holm adjustment across 8 associations.

Statistical Analysis

AUC differences were evaluated using 10,000 paired, patient-clustered percentile bootstrap replicates and 2-sided tests. External macroaverages assigned equal weights to the CHSZCAMS and WSPH validation cohorts. The 95% CIs were unadjusted; *P* values were Holm-adjusted within the specified comparison families. Comparisons using unadjusted *P* values are identified in the corresponding results and tables. Study precision was characterized retrospectively using paired-bootstrap estimates, with calculations detailed in Section S13 in Multimedia Appendix 1. For the color and test-domain controls, paired outcome-stratified bootstrap replicates resampled patients within external centers or examination-linked cases across both test domains, respectively. Their CIs and *P* values were unadjusted, and CIs were conditional on the fixed 5-model ensembles.

Results

Study Cohorts

The study included 2209 unique patients across 7 mutually exclusive cohorts from 3 institutions (Table 1). The CHCAMS source cohort comprised 1245 patients and 8476 paired WLI and Vir-NBI images. Separate CHCAMS cohorts were used for real-NBI fine-tuning, model development, and retrospective internal validation. Retrospective external validation was performed in 182 patients from CHSZCAMS and 129 patients from WSPH. The prospective CHCAMS validation cohort comprised 138 patients: 62 with benign lesions, 35 with dysplasia, and 41 with invasive cancer.

Table 1. Study cohorts and participant characteristics.

Characteristic	CHCAMS[a] source	CHCAMS fine-tuning	CHCAMS development	CHCAMS retrospective validation	CHSZCAMS[b] validation	WSPH[c] validation	CHCAMS prospective validation
Demographic characteristics							
Patients, n	1245	100	207	208	182	129	138
Age (y), mean (SD)	52.1 (8.4)	52.2 (8.1)	51.8 (8.1)	52.1 (8.0)	50.2 (8.9)	53.3 (7.4)	51.4 (7.7)
Male, n (%)	670 (53.8)	51 (51.0)	101 (48.8)	103 (49.5)	99 (54.4)	64 (49.6)	66 (47.8)
Total images, n	8476	484	1040	1014	872	644	674
Binary study end point, n (%)							
Low-risk	651 (52.3)	50 (50)	108 (52.2)	109 (52.4)	97 (53.3)	66 (51.2)	72 (52.2)
High-risk	594 (47.7)	50 (50)	99 (47.8)	99 (47.6)	85 (46.7)	63 (48.8)	66 (47.8)

Characteristic	CHCAMS[a] source	CHCAMS fine-tuning	CHCAMS development	CHCAMS retrospective validation	CHSZCAMS[b] validation	WSPH[c] validation	CHCAMS prospective validation
Pathological diagnosis, n (%)							
Inflammatory lesion	87 (6.9)	6 (6)	15 (7.2)	17 (8.2)	14 (7.7)	11 (8.5)	12 (8.7)
Hyperplasia	120 (9.6)	10 (10)	25 (12.1)	23 (11.1)	21 (11.5)	14 (10.9)	15 (10.9)
Granuloma	37 (2.9)	3 (3)	6 (2.9)	6 (2.9)	4 (2.2)	3 (2.3)	4 (2.9)
Papilloma	17 (1.4)	2 (2)	3 (1.4)	3 (1.4)	3 (1.6)	2 (1.6)	2 (1.4)
Tuberculosis	2 (0.2)	0 (0)	1 (0.5)	0 (0)	1 (0.5)	0 (0)	0 (0)
Polyp	291 (23.4)	21 (21)	42 (20.3)	42 (20.2)	37 (20.3)	26 (20.2)	28 (20.3)
Edema	19 (1.5)	1 (1)	2 (0.9)	2 (0.9)	2 (1.1)	1 (0.8)	1 (0.7)
Cyst	7 (0.6)	1 (1)	1 (0.5)	1 (0.5)	1 (0.5)	0 (0)	0 (0)
Mild dysplasia	71 (5.7)	6 (6)	13 (6.3)	15 (7.2)	14 (7.7)	9 (7.0)	10 (7.2)
Moderate dysplasia	48 (3.9)	5 (5)	8 (3.9)	10 (4.8)	9 (4.9)	7 (5.4)	6 (4.3)
Severe dysplasia	131 (10.5)	14 (14)	28 (13.5)	27 (12.9)	22 (12.1)	17 (13.2)	19 (13.8)
Invasive cancer	415 (33.3)	31 (31)	63 (30.4)	62 (29.8)	54 (29.7)	39 (30.2)	41 (29.7)

[a]CHCAMS: National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College.
[b]CHSZCAMS: Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College.
[c]WSPH: People's Hospital of Wenshan Prefecture.

Real-NBI Label Efficiency

Patient-level AUC in the CHCAMS development cohort was higher with WLI pretraining than with Vir-NBI pretraining at each real-NBI training-set size (Figure 2D). With 10 labeled patients, the mean AUC was 0.863 (95% CI 0.816-0.903) for the WLI-pretrained strategy and 0.788 (95% CI 0.729-0.841) for the Vir-NBI–pretrained strategy, corresponding to a paired difference of –0.074 (Vir-NBI minus WLI; 95% CI –0.126 to –0.025; Holm-adjusted *P*=.02). The difference narrowed progressively as the number of labeled patients increased: –0.061 at 20 patients, –0.047 at 40 patients, –0.038 at 60 patients, –0.023 at 80 patients, and –0.014 at 100 patients. After adjustment across the 6 training-set-size comparisons, the difference at 10 patients remained statistically significant (Table 2).

Integrated performance across the CHCAMS development-cohort learning curves was also lower with Vir-NBI pretraining. The normalized AULC difference was –0.058 over 10 to 40 labeled patients (95% CI –0.106 to –0.013; Holm-adjusted *P*=.02) and –0.040 over 10 to 100 labeled patients (95% CI –0.080 to –0.002; Holm-adjusted *P*=.04; Table 2).

Across the CHSZCAMS and WSPH external validation cohorts, mean patient-level AUC with 10 labeled real-NBI patients for fine-tuning was 0.843 with WLI and 0.773 with Vir-NBI, corresponding to an equal-center difference of –0.070 (95% CI –0.116 to –0.029; Holm-adjusted *P*=.006; Figure 2E-F). Differences at 20 and 40 patients were –0.067 (95% CI –0.116 to –0.023) and –0.055 (95% CI –0.100 to –0.014; Holm-adjusted *P*=.009 and *P*=.03). At 100 patients, the difference was –0.027 (95% CI –0.061 to 0.008; Holm-adjusted *P*=.13). External normalized AULC differences were –0.064 (95% CI –0.110 to –0.023) over 10 to 40 patients (Holm-adjusted *P*=.004) and –0.049 (95% CI –0.090 to –0.011) over 10 to 100 patients (Holm-adjusted *P*=.01). Estimates for all training-set sizes and individual centers are reported in Tables S9 to S12 in Multimedia Appendix 1.

Figure 2. Validation performance and learning curves for white-light imaging (WLI) and virtual narrow-band imaging (Vir-NBI) pretraining. (A-C) Image-level receiver operating characteristic (ROC) curves for the 5-model ensembles in the CHSZCAMS, WSPH, and prospective CHCAMS validation cohorts, respectively. Shading shows pointwise 95% confidence bands from 10,000 risk-stratified patient-cluster bootstrap samples. (D and E) Mean patient-level area under the ROC curve (AUCs) across 25 paired runs (5 patient-subset sequences × 5 model seeds) in the CHCAMS development cohort and the CHSZCAMS and WSPH external validation cohorts, respectively. External AUCs are equally weighted averages of the 2 centers. Shading shows pointwise 95% bootstrap CIs accounting for patient, subset-sequence, and model-seed variation. (F) Paired external AUCs at training-set sizes of 10, 40, and 100 labeled patients. Lines connect strategies sharing the same subset sequence and model seed. CHCAMS: National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College; CHSZCAMS: Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College; WSPH: People's Hospital of Wenshan Prefecture.

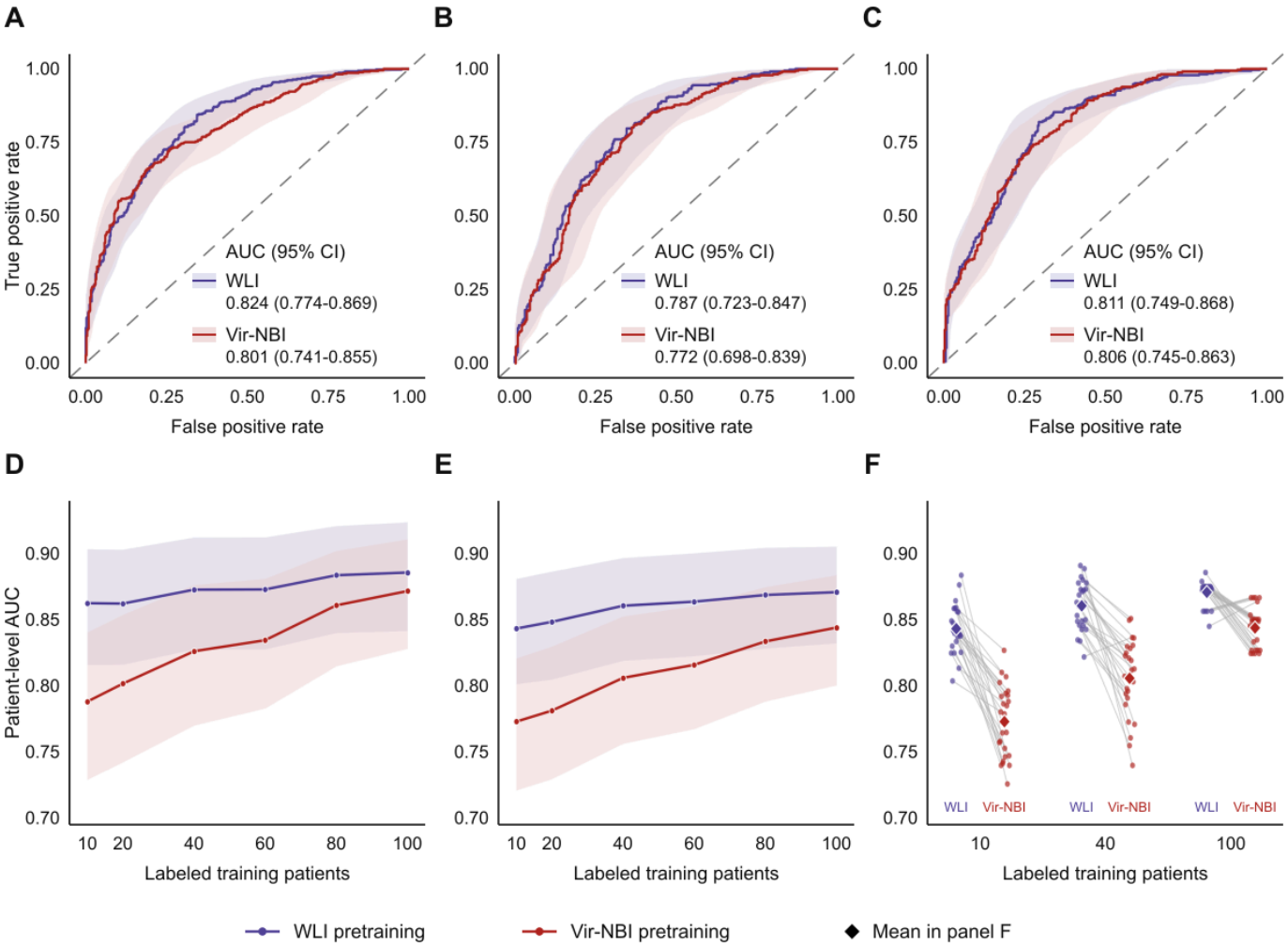


Table 2. Retrospective real-NBI label efficiency in the CHCAMS[a] development cohort.

Outcome	WLI[b] strategy	Vir-NBI[c] strategy	Difference (95% CI[d])	Adjusted *P* value
Patient-level AUC[e], 10 patients	0.863	0.788	–0.074 (–0.126 to –0.025)	.02
Patient-level AUC[f], 100 patients	0.886	0.872	–0.014 (–0.049 to 0.020)	.41
Normalized AULC, 10-40 patients	0.866	0.808	–0.058 (–0.106 to –0.013)	.02[g]
Normalized AULC, 10-100 patients	0.874	0.835	–0.040 (–0.080 to –0.002)	.04

[a]CHCAMS: National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College.
[b]WLI: white-light imaging.
[c]Vir-NBI: virtual narrow-band imaging.
[d]All 95% CIs are unadjusted.
[e]AUC: area under the receiver operating characteristic curve.
[f]AULC: area under the learning curve.
[g]Holm correction covers 6 training-set-size comparisons and, separately, 2 AULCs.

Retrospective Multicenter Validation

In the CHSZCAMS external cohort, image-level AUC was 0.824 (95% CI 0.774-0.869) with WLI pretraining and 0.801 (95% CI 0.741-0.855) with Vir-NBI pretraining. The corresponding AUCs in the WSPH external cohort were 0.787 (95% CI 0.723-0.847) and 0.772 (95% CI 0.698-0.839), respectively (Figure 2A-B; Table 3). The equal-weight external macroaverage was 0.806 for the WLI-pretrained strategy and 0.787 for the Vir-NBI–pretrained strategy, with a

difference of –0.019 (95% CI –0.047 to 0.008; *P*=.17; Table 3). Retrospective precision calculations for the CHSZCAMS and WSPH external validation cohorts gave an approximate minimum detectable AUC difference of 0.040 at 80% power (2-sided α=.05). This estimate does not establish equivalence between the strategies. Precision estimates for each learning-curve training-set size are reported in Table S17 in Multimedia Appendix 1.

Table 3. Retrospective external validation in CHSZCAMS[a] and WSPH[b].

Outcome	WLI[c] strategy	Vir-NBI[d] strategy	Difference (95% CI)[e]	*P* value
Image-level AUC[f]	0.806	0.787	–0.019 (–0.047 to 0.008)	.17[g]
Patient-level AUC	0.900	0.869	–0.032 (–0.060 to –0.005)	.053[h]

[a]CHSZCAMS: Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College.
[b]WSPH: People's Hospital of Wenshan Prefecture.
[c]WLI: white-light imaging.
[d]Vir-NBI: virtual narrow-band imaging.
[e]All 95% CIs are unadjusted.
[f]AUC: area under the receiver operating characteristic curve.
[g]Unadjusted primary comparison.
[h]*P* values are adjusted using the Holm method across the 5 factorial contrasts reported in Table 4, including the external patient-level AUC comparison of Vir-NBI vs WLI pretraining after real-NBI fine-tuning.

At the patient level, AUC was 0.884 with WLI pretraining and 0.840 with Vir-NBI pretraining in the CHCAMS retrospective internal validation cohort. Patient-level AUCs were 0.900 and 0.881, respectively, in the CHSZCAMS external validation cohort and 0.901 and 0.856 in the WSPH external validation cohort. Across the 2 external cohorts, the patient-level macroaverage difference was –0.032 (95% CI –0.060 to –0.005; Holm-adjusted *P*=.053).

Effects of Virtual Translation and Real-NBI Fine-Tuning

The 2×2 analysis in the CHSZCAMS and WSPH external validation cohorts examined the effects of replacing source WLI with Vir-NBI before and after real-NBI fine-tuning (Figure S3 in Multimedia Appendix 1). In the absence of real-NBI fine-tuning, patient-level external macroaverage AUC was 0.076 lower with Vir-NBI direct transfer than with WLI direct transfer (95% CI –0.119 to –0.033; Holm-adjusted *P*=.002). After real-NBI fine-tuning, the corresponding difference between the Vir-NBI–pretrained and WLI-pretrained strategies was –0.032 (95% CI –0.060 to –0.005; Holm-adjusted *P*=.053; Table 4). Real-NBI fine-tuning increased AUC by 0.034 (95% CI 0.001-0.068; Holm-adjusted *P*=.08) within the WLI pathway and by 0.079 (95% CI 0.041-0.117; Holm-adjusted *P*<.001) within the Vir-NBI pathway. The interaction estimate was 0.044 (95% CI 0.002-0.087; Holm-adjusted *P*=.08) and was not statistically significant after Holm adjustment.

Table 4. Retrospective factorial analysis in the CHSZCAMS[a] and WSPH[b] external validation cohorts.

Contrast	Difference in patient-level AUC[c] (95% CI)[d]	Adjusted *P* value
Vir-NBI[e] vs WLI[f] without real-NBI[g] fine-tuning	–0.076 (–0.119 to –0.033)	.002
Vir-NBI vs WLI after real-NBI fine-tuning	–0.032 (–0.060 to –0.005)	.053[h]
Real-NBI fine-tuning after WLI pretraining	0.034 (0.001 to 0.068)	.08
Real-NBI fine-tuning after Vir-NBI pretraining	0.079 (0.041 to 0.117)	<.001
Interaction	0.044 (0.002 to 0.087)	.08

[a]CHSZCAMS: Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College.
[b]WSPH: People's Hospital of Wenshan Prefecture.
[c]AUC: area under the receiver operating characteristic curve.
[d]All 95% CIs are unadjusted.
[e]Vir-NBI: virtual narrow-band imaging.
[f]WLI: white-light imaging.
[g]NBI: narrow-band imaging.
[h]*P* values are adjusted using the Holm method across the 5 factorial contrasts, including the external patient-level AUC comparison of Vir-NBI vs WLI pretraining after real-NBI fine-tuning.

Strongly color-augmented WLI pretraining yielded an equal-weighted patient-level AUC across the CHSZCAMS and WSPH external validation cohorts of 0.899 vs 0.900 for standard WLI (difference –0.001, 95% CI –0.020 to 0.018; Holm-adjusted *P*>.99). Image-level AUC was 0.803 vs 0.806 (difference –0.003, 95% CI –0.020 to 0.014; Holm-adjusted *P*>.99).

For models trained for 10 source-pretraining and 10 fine-tuning epochs, the equal-weight patient-level AUC across the CHSZCAMS and WSPH external validation cohorts was 0.876 with WLI and 0.830 with Vir-NBI, using 5-seed ensembles (difference –0.046, 95% CI –0.076 to –0.017; unadjusted *P*=.002). The image-level difference was –0.045 (95% CI –0.075 to –0.016; unadjusted *P*=.003).

The WLI color-control arm achieved a mean ΔE76 of 29.412 (SD 4.9) in the verification sample, compared with 29.363 (SD 5.5) for the translation arm. Under the same 10+10 epoch schedule, its equal-weight image-level AUC across the CHSZCAMS and WSPH external validation cohorts was 0.774, vs 0.784 for WLI and 0.739 for Vir-NBI. The color-control differences were –0.010 vs WLI (95% CI –0.031 to 0.012) and 0.036 vs Vir-NBI (95% CI 0.004-0.070). Corresponding patient-level AUCs were 0.873, 0.876, and 0.830; the color-control differences were –0.003 vs WLI (95% CI –0.023 to 0.016) and 0.043 vs Vir-NBI (95% CI 0.014-0.074). CIs are unadjusted. The color-control arm had higher image- and patient-level AUCs than Vir-NBI in all 5 seeds (Figure 3B; Tables S13-S14 in Multimedia Appendix 1).

Figure 3. Clinician's ratings and analyses of color and test modality. (A) Virtual narrow-band imaging (Vir-NBI) score distributions for 100 CHCAMS source-cohort image pairs assessed by 12 clinicians (1200 ratings/item); scores range from 1 to 5, with higher scores indicating better quality. (B) Image-level area under the receiver operating characteristic curve (AUCs) in the CHSZCAMS and WSPH external validation cohorts after 10 pretraining and 10 fine-tuning epochs. Color augmentation matched the mean white-light imaging (WLI)–to–Vir-NBI color change. Connected circles denote matched seeds; black bars denote ensemble AUCs with 95% bootstrap CIs. (C and D) Primary-model case-level receiver operating characteristic (ROC) curves on real narrow-band imaging (real-NBI) and translated close-view WLI, respectively, in the same 58 examination-linked cases. Shading denotes pointwise 95% bootstrap confidence bands. All CIs are unadjusted. CP: clinical plausibility; CHCAMS: National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College; CHSZCAMS: Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College; DC: diagnostic confidence; FBDO: freedom from boundary distortion/oversmoothing; FVA: freedom from vascular artifacts; WSPH: People's Hospital of Wenshan Prefecture.

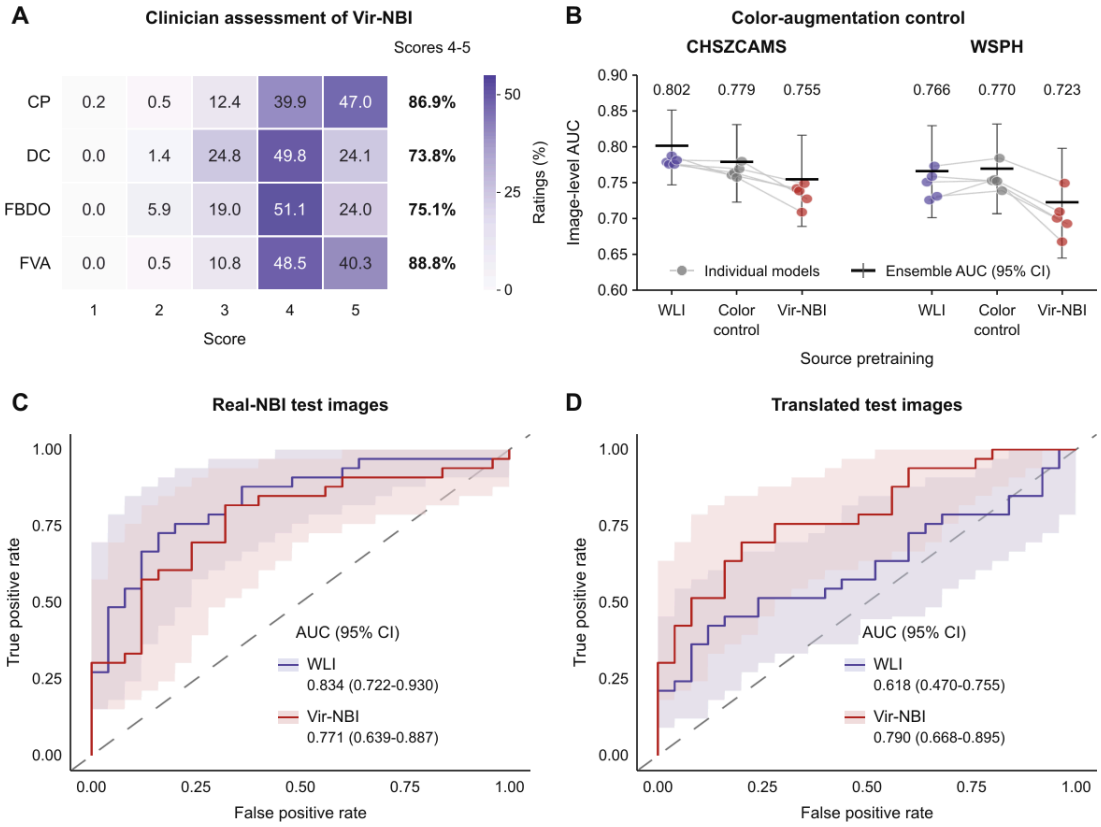


Image Appearance and Downstream Utility

Feature-space relationships between WLI, Vir-NBI, and real-NBI varied according to the encoder. In ImageNet feature space, the RFF MMD² (random Fourier feature maximum mean discrepancy squared) difference for distance to real-NBI was 0.165 higher for Vir-NBI than for WLI (95% CI 0.156-0.175), indicating that translation increased rather than decreased the measured distance. The estimate was also positive with the WLI-pretrained encoder but negative with the Vir-NBI–pretrained encoder (Table S6A and Figure S6a in Multimedia Appendix 1).

In the close-view analysis of the 58-case examination-linked test-modality cohort, the 10+10 epoch WLI- and Vir-NBI–pretrained models had case-level AUCs of 0.792 and 0.687 on real-NBI and 0.676 and 0.736 on translated Vir-NBI. The test-domain interaction was 0.164 (95% CI –0.008 to 0.330). Models used in the primary analysis showed the same direction: AUCs were 0.834 (95% CI 0.722-0.930) and 0.771 (95% CI 0.639-0.887) on real-NBI and 0.618 (95% CI 0.470-0.755) and 0.790 (95% CI 0.668-0.895) on translated Vir-NBI (Figure 3C-D), with an interaction of 0.235 (95% CI 0.080-0.390). Close-view case-level interactions were positive for all 5 seeds under both model settings. When all WLI views were included, case-level interactions were 0.103 (95% CI –0.063 to 0.270) and 0.144 (95% CI –0.007 to 0.297), respectively. These

pointwise intervals are unadjusted; both model settings were evaluated in the same cases (Tables S15-S16 in Multimedia Appendix 1).

In the CHSZCAMS and WSPH external validation cohorts, the patient-level linear-probe AUC difference between Vir-NBI and WLI representations was –0.056 (95% CI –0.118 to 0.005; *P*=.07). Translation was associated with a decrease of 0.099 in lesion-region attribution mass (95% CI –0.106 to –0.093) and a decrease of 0.012 in attribution-centroid distance (95% CI –0.015 to –0.009; Table S6 and Figure S6 in Multimedia Appendix 1).

The 12-clinician assessment comprised 100 image pairs from the CHCAMS source cohort and 4800 item scores. Mean scores were 4.33 (SD 0.72) for clinical plausibility, 3.97 (SD 0.74) for diagnostic confidence, 3.93 (SD 0.81) for freedom from boundary distortion/oversmoothing, and 4.29 (SD 0.67) for freedom from vascular artifacts (Table S7A in Multimedia Appendix 1). Scores of 4 to 5 occurred in 86.9% (1043/1200), 73.8% (886/1200), 75.1% (901/1200), and 88.8% (1065/1200) of ratings, respectively (Figure 3A; Table S7D in Multimedia Appendix 1). Clinical plausibility ratings were concentrated at the upper end of the scale, with 47% (564/1200) of ratings at the maximum score. Single-clinician absolute-agreement ICCs ranged from 0.052 to 0.104; ICCs for the mean of 12 clinicians ranged from 0.398 (95% CI 0.265-0.497) to 0.583 (95% CI 0.471-0.659; Table S7B in Multimedia Appendix 1). None of the 8 score-appearance or score-utility associations was statistically significant after Holm correction (Table S7C in Multimedia Appendix 1).

Evaluation of the Prospectively Collected CHCAMS Cohort

The prospective CHCAMS cohort included 138 patients with 674 real-NBI images.

Image-level AUC was 0.811 (95% CI 0.749-0.868) with WLI pretraining and 0.806 (95% CI 0.745-0.863) with Vir-NBI pretraining, corresponding to a difference of –0.005 (95% CI –0.038 to 0.028; *P*=.78; Figure 2C). Patient-level AUC was 0.846 (95% CI 0.775-0.910) and 0.844 (95% CI 0.777-0.905), respectively, with a difference of –0.001 (95% CI –0.047 to 0.044; *P*=.96; Table 5).

Sensitivity was 90.9% (60/66) with WLI and 81.8% (54/66) with Vir-NBI, an absolute difference of –9.1 percentage points (unadjusted 95% bootstrap CI –16.7 to –3.0; Holm-adjusted bootstrap *P*=.007). Specificity was 72.2% (52/72) vs 68.1% (49/72), a difference of –4.2 percentage points (95% CI –11.1 to 2.8; Holm-adjusted bootstrap *P*=.33).

Among the 66 high-risk patients, 54 were correctly classified by both strategies, 6 only by WLI, none only by Vir-NBI, and 6 by neither. The exact McNemar *P* value was .03 and .06 after Holm correction across the 2 operating-point tests. Among the 72 low-risk patients, the corresponding counts were 47, 5, 2, and 18 (exact *P*=.45).

For uncalibrated patient-level probabilities, the calibration intercept, slope, and expected calibration error were 0.039, 0.831, and 0.101, respectively, for the WLI-pretrained strategy and –0.189, 1.292, and 0.083 for the Vir-NBI–pretrained strategy (Table 6).

Table 5. Performance in the prospective CHCAMS[a] validation cohort: discrimination and operating characteristics.

Outcome	WLI[b] strategy (95% CI)[c]	Vir-NBI[d] strategy (95% CI)	Difference (95% CI)	*P* value[e]
Image-level AUC[f]	0.811 (0.749-0.868)	0.806 (0.745-0.863)	–0.005 (–0.038 to 0.028)	.78[g]
Patient-level AUC	0.846 (0.775-0.910)	0.844 (0.777-0.905)	–0.001 (–0.047 to 0.044)	.96[g]
Sensitivity, n/N (%)	60/66 (90.9)	54/66 (81.8)	–0.091 (–0.167 to –0.030)	.007[h]
Specificity, n/N (%)	52/72 (72.2)	49/72 (68.1)	–0.042 (–0.111 to 0.028)	.33[h]

[a]CHCAMS: National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College.
[b]WLI: white-light imaging.
[c]CIs are unadjusted pointwise intervals.
[d]Vir-NBI: virtual narrow-band imaging.
[e]*P* values for sensitivity and specificity.
[f]AUC: area under the receiver operating characteristic curve.
[g]Unadjusted AUC tests.
[h]Holm-adjusted bootstrap.

Table 6. Performance in the prospective CHCAMS[a] validation cohort: calibration and overall predictive accuracy.[b]

Outcome	WLI[c] strategy (95% CI)[b]	Vir-NBI[d] strategy (95% CI)[b]
Calibration intercept	0.039 (–0.201 to 0.287)	–0.189 (–0.499 to 0.064)
Calibration slope	0.831 (0.525 to 1.395)	1.292 (0.916 to 1.954)

Outcome	WLI[c] strategy (95% CI)[b]	Vir-NBI[d] strategy (95% CI)[b]
Expected calibration error	0.101 (0.086 to 0.184)	0.083 (0.066 to 0.160)
Brier score	0.157 (0.122 to 0.194)	0.164 (0.135 to 0.194)
Log loss	0.505 (0.397 to 0.628)	0.494 (0.426 to 0.565)

[a]CHCAMS: National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College.
[b]CIs are unadjusted pointwise intervals.
[c]WLI: white-light imaging.
[d]Vir-NBI: virtual narrow-band imaging.

Discussion

Principal Findings

The primary comparison across the CHSZCAMS and WSPH external validation cohorts did not demonstrate incremental discrimination from Vir-NBI pretraining, with a CI that included both lower performance and a small benefit. Learning curves in both external cohorts favored WLI when few labeled real-NBI patients were available, extending the pattern observed in the CHCAMS development cohort. WLI also had higher external AUCs when both pathways used fixed training durations. Transforming WLI into an NBI-like appearance did not yield a demonstrated increase in training value in the evaluated framework.

The factorial estimates were consistent with improved performance after fine-tuning on labeled real-NBI, particularly within the Vir-NBI pathway. However, the interaction was not statistically significant after multiplicity adjustment, leaving uncertainty about whether the magnitude of the fine-tuning effect differed between pathways.

The cross-analysis accounted for variation in patient selection and model training. The AULC estimates for the CHSZCAMS and WSPH external validation cohorts summarized the advantage of WLI across the evaluated training-set sizes, while the smaller difference at 100 labeled patients had a CI that included 0. This pattern suggests that the relative value of source pretraining depends on the amount of labeled target-modality data and supports evaluating pretraining strategies across different numbers of labeled real-NBI patients.

Interpretation in the Context of Synthetic Medical Imaging

A deterministic translation generates a new visual representation but does not add new patients, pathology labels, or independently acquired target-modality information. Any training benefit must therefore arise from transforming information already present in WLI into a representation that is more compatible with real-NBI. In principle, such a transformation could emphasize task-relevant vascular or mucosal features. Conversely, it could suppress subtle lesion information, alter texture, or introduce generator-specific patterns that are visually plausible but unhelpful for classification. Synthetic medical imaging can address some forms of data scarcity [16,19], but visual fidelity and apparent domain similarity do not guarantee the preservation of task-relevant information [20,21].

This distinction determines the appropriate comparator. Comparing Vir-NBI pretraining only with ImageNet initialization or training from limited real-NBI data would test whether source-domain supervision is useful but would not establish the incremental value of translation. Because Vir-NBI images were deterministically generated from pathology-labeled WLI, direct pretraining on the same WLI images provided the relevant comparison, with source patients, labels, architecture, learning rate, optimizer, and target-modality data held constant.

The color controls and test-domain controls suggest that the mean color-change magnitude alone is insufficient to explain the weaker Vir-NBI results. Performance after color augmentation did not differ significantly from WLI, although this does not establish equivalence. The shift in relative performance across test domains, together with the larger ImageNet-space distance from Vir-NBI to real-NBI, supports generator-specific domain mismatch as a plausible contributor. Interpretation remains limited by encoder dependence in the distance estimates, unmatched color direction, texture, and geometry, and imprecise case-level interactions in the all-view analyses. The results are compatible with multiple mechanisms and can neither distinguish domain mismatch from loss of diagnostic information nor establish a clinical benefit with real-NBI.

Because downstream usefulness also depends on calibration and clinical operating characteristics, discrimination alone is insufficient [35,36]. Prior laryngeal image-classification studies have likewise reported variability across model architectures and datasets [37-39].

Clinical optical-imaging studies address a different question. Filipovský et al [40] prospectively compared preoperative and perioperative WLI/NBI assessments against histopathology in 87 patients, supporting NBI-guided diagnosis and biopsy rather than demonstrating the efficacy of virtual-image pretraining. Hu et al [41] studied 142 patients with suspected laryngeal cancer using selectively available optical-biopsy modalities, including NBI, and retrospectively applied AI to prospectively collected endoscopic images. Holý et al [42] evaluated ORBEYE exoscopy with WLI and NBI in 84 patients and described targeted perioperative biopsy. These studies support clinical investigation of acquired optical information; they do not establish that

deterministic WLI translation improves classifier learning on real-NBI.

The clinician's ratings of images from the CHCAMS source cohort supported the visual plausibility of Vir-NBI and indicated generally few perceived artifacts. Favorable visual quality, however, was not accompanied by a demonstrated improvement in real-NBI classification.

Clinical and Methodological Implications

These findings do not question the clinical value of optical NBI. Optical NBI is an acquired imaging modality that reveals vascular and mucosal features that may be less apparent under conventional WLI [1-3]. Vir-NBI is a deterministic transformation of an existing WLI frame and cannot recover information that was not captured in the source image. Acquired NBI and virtual target-style translation should therefore not be regarded as interchangeable sources of training information.

The CHCAMS prospective validation cohort extended the evaluation to patients recruited after model development. Patient-level AUC was 0.846 (95% CI 0.775-0.910) with WLI and 0.844 (95% CI 0.777-0.905) with Vir-NBI, while WLI identified 90.9% of high-risk patients compared with 81.8% for Vir-NBI. The sensitivity difference was based on 6 discordant patients and requires validation in larger cohorts. Calibration, threshold transportability, and patient-level discordance remain relevant measures of clinical performance [35,36].

The broader methodological implication is that synthetic data should be evaluated according to the clinical problem they are intended to solve. When the proposed benefit is reduced target-domain labeling, the primary evidence should come from learning curves constructed across different numbers of labeled patients. When synthetic images are derived from existing labeled source images, their incremental value should be measured against those same source images rather than against an untrained or generic-pretrained baseline. External and prospective validation should then determine whether any observed advantage persists across institutions and time. Increasing the number of image files is not equivalent to increasing the effective amount of independent training information.

Strengths and Limitations

Strengths include matching source patients and pathology labels, patient-level classifier splits, paired, clustered inference, and evaluation in the CHSZCAMS and WSPH external validation cohorts. External learning curves extended the assessment of label efficiency beyond the CHCAMS development cohort, and fixed-epoch analyses assessed sensitivity to training duration. Together, these analyses provided context for the primary comparison.

Several limitations warrant consideration. The study evaluated 1 frozen generator, 1 backbone, and laryngeal risk classification. CHCAMS development-cohort learning curves may be optimistic because the same cohort supported model selection. The color control matched only the mean color difference. The translated-test analysis included 58 examination-linked cases with nonregistered frames and limited precision; both model settings were evaluated in the same cases. Feature-space and saliency measures were descriptive and could not establish a causal mechanism, causal lesion use, or diagnostic fidelity. Source ROI polygons were reused for WLI and translated-image attribution without separate, blinded ROI annotation. Interclinician agreement was limited, and the relationship between subjective image quality and training benefit remains uncertain. The CHCAMS prospective validation cohort of 138 patients was convenience-sampled, and the small number of discordant high-risk classifications limited precision for the sensitivity difference.

Conclusions

In this multicenter comparison, Vir-NBI pretraining did not demonstrate incremental value over pretraining on the original WLI. Learning curves in the CHSZCAMS and WSPH external validation cohorts favored WLI when few labeled real-NBI patients were available. The color and test-domain analyses suggest that compatibility with the real-NBI test domain may contribute to the performance differences.

Acknowledgments

The authors thank all participating centers and endoscopists for their contributions to data collection, annotation, and reader-study evaluation.

Generative AI Use

During the preparation of this manuscript, the authors used ChatGPT (GPT-5.6) to improve the English language, grammar, and readability. All AI-assisted output was reviewed, edited, and approved by the authors, who take full responsibility for the final content of the manuscript.

Funding

This work was supported by National High-Level Hospital Clinical Research Funding (grant number LC2024A04), the CAMS Innovation Fund for Medical Sciences (CIFMS; grant number 2022-I2M-C&T-B-059), Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College (grant number E010324014), and the People's Hospital of Wenshan Prefecture (grant number ZDXM2024002).

Data Availability

Deidentified datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request, subject to institutional approval and applicable privacy restrictions.
The primary analysis code is referenced by [43]. Scripts and aggregate results for the supplementary analyses are available from the corresponding author upon reasonable request. Access to deidentified endoscopic images and patient-level data is subject to institutional approval and applicable consent and privacy requirements.

Authors' Contributions

Conceptualization: JLH, XGN
Data curation: JLH, LJL, MLW, LZD, YML
Formal analysis: JLH
Funding acquisition: XGN
Investigation: JLH, LJL, MLW, JQZ, LZD, YML, YDZ, XZ
Methodology: JLH, JQZ
Project administration: XGN
Resources: YML, XGN
Software: JLH, JQZ
Supervision: XGN
Validation: LJL, MLW, JQZ, LZD, YDZ, XZ
Visualization: JLH
Writing – original draft: JLH
Writing – review & editing: JLH, LJL, MLW, JQZ, LZD, YML, YDZ, XZ, XGN

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary methods, results, tables, and figures.
[DOCX File (Microsoft Word File), 1809 KB-Multimedia Appendix 1]

Checklist 1

TRIPOD+AI checklist.
[DOCX File (Microsoft Word File), 34 KB-Checklist 1]

Checklist 2

CLAIM 2024 checklist.
[DOCX File (Microsoft Word File), 32 KB-Checklist 2]

Checklist 3

STARD-AI checklist.
[DOCX File (Microsoft Word File), 33 KB-Checklist 3]

References

1. Sun C, Han X, Li X, Zhang Y, Du X. Diagnostic performance of narrow band imaging for laryngeal cancer: a systematic review and meta-analysis. Otolaryngol Head Neck Surg. Apr 2017;156(4):589-597. [doi: 10.1177/0194599816685701] [Medline: 28168890]
2. Chen J, Li Z, Wu T, Chen X. Accuracy of narrow-band imaging for diagnosing malignant transformation of vocal cord leukoplakia: a systematic review and meta-analysis. Laryngoscope Investig Otolaryngol. Apr 2023;8(2):508-517. [doi: 10.1002/lio2.1049] [Medline: 37090880]
3. Ahmadzada S, Vasan K, Sritharan N, et al. Utility of narrowband imaging in the diagnosis of laryngeal leukoplakia: systematic review and meta-analysis. Head Neck. Nov 2020;42(11):3427-3437. [doi: 10.1002/hed.26428] [Medline: 32965766]
4. Davaris N, Voigt-Zimmermann S, Kropf S, Arens C. Flexible transnasal endoscopy with white light or narrow band imaging for the diagnosis of laryngeal malignancy: diagnostic value, observer variability and influence of previous laryngeal surgery. Eur Arch Otorhinolaryngol. Feb 2019;276(2):459-466. [doi: 10.1007/s00405-018-5256-1] [Medline: 30569190]
5. Ni XG, Zhu JQ, Zhang QQ, Zhang BG, Wang GQ. Diagnosis of vocal cord leukoplakia: the role of a novel narrow band imaging endoscopic classification. Laryngoscope. Feb 2019;129(2):429-434. [doi: 10.1002/lary.27346] [Medline: 30229933]

6. Zhang BG, Zhu JQ, Zhang W, Su FX, Wang GQ, Ni XG. Effect of a training course on the diagnosis of vocal fold leukoplakia by narrow-band imaging. J Laryngol Otol. Oct 23, 2020;134(10):1-6. [doi: 10.1017/S002221512000211X] [Medline: 33092654]
7. Pietruszewska W, Morawska J, Rosiak O, Leduchowska A, Klimza H, Wierzbicka M. Vocal fold leukoplakia: which of the classifications of white light and narrow band imaging most accurately predicts laryngeal cancer transformation? Proposition for a diagnostic algorithm. Cancers (Basel). Jun 30, 2021;13(13):3273. [doi: 10.3390/cancers13133273] [Medline: 34208811]
8. Campo F, Ralli M, Di Stadio A, Greco A, Pellini R, de Vincentiis M. Role of narrow band imaging endoscopy in preoperative evaluation of laryngeal leukoplakia: a review of the literature. Ear Nose Throat J. Nov 2022;101(9):NP403-NP408. [doi: 10.1177/0145561320973770] [Medline: 33213196]
9. Hu R, Zhong Q, Xu ZG, et al. Application of deep convolutional neural networks in the diagnosis of laryngeal squamous cell carcinoma based on narrow band imaging endoscopy. Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi. May 7, 2021;56(5):454-458. [doi: 10.3760/cma.j.cn115330-20200927-00773] [Medline: 34010998]
10. He Y, Cheng Y, Huang Z, et al. A deep convolutional neural network-based method for laryngeal squamous cell carcinoma diagnosis. Ann Transl Med. Dec 2021;9(24):1797. [doi: 10.21037/atm-21-6458] [Medline: 35071491]
11. Esmaeili N, Sharaf E, Gomes Ataide EJ, et al. Deep convolution neural network for laryngeal cancer classification on contact endoscopy-narrow band imaging. Sensors (Basel). Dec 6, 2021;21(23):8157. [doi: 10.3390/s21238157] [Medline: 34884166]
12. Azam MA, Sampieri C, Ioppi A, et al. Deep learning applied to white light and narrow band imaging videolaryngoscopy: toward real-time laryngeal cancer detection. Laryngoscope. Sep 2022;132(9):1798-1806. [doi: 10.1002/lary.29960] [Medline: 34821396]
13. Sampieri C, Azam MA, Ioppi A, et al. Real-time laryngeal cancer boundaries delineation on white light and narrow-band imaging laryngoscopy with deep learning. Laryngoscope. Jun 2024;134(6):2826-2834. [doi: 10.1002/lary.31255] [Medline: 38174772]
14. Huang JL, Zhu JQ, Zhao YD, et al. ROI-guided virtual narrow band imaging for laryngeal cancer screening. iScience. Jun 19, 2026;29(6):116161. [doi: 10.1016/j.isci.2026.116161] [Medline: 42256307]
15. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); Oct 22-29, 2017; Venice. [doi: 10.1109/ICCV.2017.244]
16. Koshino K, Werner RA, Pomper MG, et al. Narrative review of generative adversarial networks in medical and molecular imaging. Ann Transl Med. May 2021;9(9):821. [doi: 10.21037/atm-20-6325] [Medline: 34268434]
17. Yu B, Wang Y, Wang L, Shen D, Zhou L. Medical image synthesis via deep learning. Adv Exp Med Biol. 2020;1213:23-44. [doi: 10.1007/978-3-030-33128-3_2] [Medline: 32030661]
18. Guan H, Liu M. Domain adaptation for medical image analysis: a survey. IEEE Trans Biomed Eng. Mar 2022;69(3):1173-1185. [doi: 10.1109/TBME.2021.3117407] [Medline: 34606445]
19. Koetzier LR, Wu J, Mastrodicasa D, et al. Generating synthetic data for medical imaging. Radiology. Sep 2024;312(3):e232471. [doi: 10.1148/radiol.232471] [Medline: 39254456]
20. Kazeminia S, Baur C, Kuijper A, et al. GANs for medical image analysis. Artif Intell Med. Sep 2020;109. [doi: 10.1016/j.artmed.2020.101938] [Medline: 34756215]
21. Ibrahim M, Khalil YA, Amirrajab S, et al. Generative AI for synthetic data across multiple medical modalities: a systematic review of recent developments and challenges. Comput Biol Med. May 2025;189. [doi: 10.1016/j.compbiomed.2025.109834] [Medline: 40023073]
22. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell. May 2022;4(3):e210064. [doi: 10.1148/ryai.210064] [Medline: 35652114]
23. Kelly BS, Judge C, Bollard SM, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol. Nov 2022;32(11):7998-8007. [doi: 10.1007/s00330-022-08784-6] [Medline: 35420305]
24. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. Nov 2018;15(11):e1002683. [doi: 10.1371/journal.pmed.1002683] [Medline: 30399157]
25. Bensoussan Y, Vanstrum EB, Johns MM, Rameau A. Artificial intelligence and laryngeal cancer: from screening to prognosis: a state-of-the-art review. Otolaryngol Head Neck Surg. Mar 2023;168(3):319-329. [doi: 10.1177/01945998221110839] [Medline: 35787073]
26. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. Jan 15, 2024;384:e074820. [doi: 10.1136/bmj-2023-074820] [Medline: 38224968]

27. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
28. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell. Jul 2024;6(4):e240300. [doi: 10.1148/ryai.240300] [Medline: 38809149]
29. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. Oct 2025;31(10):3283-3289. [doi: 10.1038/s41591-025-03953-8] [Medline: 40954311]
30. WHO Classification of Tumours Editorial Board. Head and Neck Tumours. 5th ed. International Agency for Research on Cancer; 2024. ISBN: 9789283245148
31. Zidar N, Gale N. Update from the 5th edition of the World Health Organization classification of head and neck tumors: hypopharynx, larynx, trachea and parapharyngeal space. Head Neck Pathol. Mar 2022;16(1):31-39. [doi: 10.1007/s12105-021-01405-6] [Medline: 35312977]
32. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016; Las Vegas, NV, USA. [doi: 10.1109/CVPR.2016.90]
33. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. May 24, 2017;60(6):84-90. [doi: 10.1145/3065386]
34. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: ICML'20: Proceedings of the 37th International Conference on Machine Learning. JMLR; 2020:1597-1607. [doi: 10.5555/3524938.3525087]
35. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc. Apr 1, 2020;27(4):621-633. [doi: 10.1093/jamia/ocz228] [Medline: 32106284]
36. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. Mar 2015;68(3):279-289. [doi: 10.1016/j.jclinepi.2014.06.018] [Medline: 25179855]
37. Xiong H, Lin P, Yu JG, et al. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. EBioMedicine. Oct 2019;48:92-99. [doi: 10.1016/j.ebiom.2019.08.075] [Medline: 31594753]
38. Xu ZH, Fan DG, Huang JQ, Wang JW, Wang Y, Li YZ. Computer-aided diagnosis of laryngeal cancer based on deep learning with laryngoscopic images. Diagnostics (Basel). Dec 14, 2023;13(24):3669. [doi: 10.3390/diagnostics13243669] [Medline: 38132254]
39. Ren J, Jing X, Wang J, et al. Automatic recognition of laryngoscopic images using a deep-learning technique. Laryngoscope. Nov 2020;130(11):E686-E693. [doi: 10.1002/lary.28539] [Medline: 32068890]
40. Filipovsky T, Kalfert D, Lukavcova E, et al. The importance of preoperative and perioperative narrow band imaging endoscopy in the diagnosis of pre-tumor and tumor lesions of the larynx. J Appl Biomed. Sep 2023;21(3):107-112. [doi: 10.32725/jab.2023.015] [Medline: 37747310]
41. Hu R, Liu X, Zhang Y, Arthur C, Qin D. Comparison of clinical nasal endoscopy, optical biopsy, and artificial intelligence in early diagnosis and treatment planning in laryngeal cancer: a prospective observational study. Front Oncol. 2025;15:1582011. [doi: 10.3389/fonc.2025.1582011] [Medline: 40556680]
42. Holy R, Filipovsky T, Lukavcova E, et al. Applying the technology ORBEYE^{TM} exoscope in transoral exoscopic laryngeal surgery: single centre prospective study. J Appl Biomed. Sep 2025;23(3):138-143. [doi: 10.32725/jab.2025.010] [Medline: 41026948]
43. Code, statistical analysis plan and model weights for "Clinical Plausibility and Pretraining Utility of Virtual Narrow-Band Imaging for Laryngeal Lesion Classification: Multicenter Model Development and Validation Study". Zenodo. 2026. URL: https://doi.org/10.5281/zenodo.21870549 [Accessed 2026-10-30]

Abbreviations

AUC: area under the receiver operating characteristic curve
AULC: area under the learning curve
CHCAMS: National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
CHSZCAMS: Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
CLAIM: Checklist for Artificial Intelligence in Medical Imaging
CycleGAN: cycle-consistent generative adversarial network
ICC: intraclass correlation coefficient

NBI: narrow-band imaging
RFF MMD²: random Fourier feature maximum mean discrepancy squared
ROI: region of interest
STARD-AI: Standards for Reporting Diagnostic Accuracy–AI
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI
Vir-NBI: virtual narrow-band imaging
WLI: white-light imaging
WSPH: People's Hospital of Wenshan Prefecture

Edited by Ivan Steenstra; peer-reviewed by Jun Zhang, Richard Holy; submitted 17.Aug.2026; final revised version received 15.Oct.2026; accepted 16.Oct.2026; published 26.Nov.2026

Please cite as:
Huang JL, Li LJ, Wang ML, Zhu JQ, Dou LZ, Liu YM, Zhao YD, Zhang X, Ni XG
Clinical Plausibility and Pretraining Utility of Virtual Narrow-Band Imaging for Laryngeal Lesion Classification: Multicenter Model Development and Validation Study
JMIR Med Inform 2026;14:e109811
URL: https://medinform.jmir.org/2026/1/e109811
doi: 10.2196/109811

© Jie-Lin Huang, Li-Juan Li, Mei-Ling Wang, Ji-Qing Zhu, Li-Zhou Dou, Yu-Meng Liu, Yu-Da Zhao, Xue Zhang, Xiao-Guang Ni. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 26.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.