

A National Multicenter Study on Artificial Intelligence-Based Detection of Nasopharyngeal Carcinoma from Endoscopic Images

Kristian Eriksen^{1*}, Solveig Berg¹, Magnus Dahl²

¹ Department of Head and Neck Oncology and AI Diagnostics, Faculty of Medicine, University of Oslo, Oslo, Norway.

² Department of Nasopharyngeal Carcinoma Research, Faculty of Medicine, University of Bergen, Bergen, Norway.

*E-mail ✉ kristian.eriksen@medisin.uio.no

Abstract

Early detection of nasopharyngeal carcinoma greatly increases the chances of successful treatment. However, the nasopharynx's concealed position and the visual overlap of its lesions with those of various benign disorders frequently complicate accurate identification, leading to delayed or missed diagnoses. This project aimed to build a deep learning tool to help otolaryngologists distinguish nasopharyngeal carcinoma, benign hyperplasia, and healthy nasopharyngeal tissue during routine endoscopic procedures. Within this nationwide, multi-institutional effort focused on model creation and testing, researchers designed the Swin Transformer-based Nasopharyngeal Diagnostic (STND) framework to categorize endoscopic views into nasopharyngeal carcinoma, benign hyperplasia, or normal tissue. Training relied on 27,362 endoscopic images (including 10,693 biopsy-verified nasopharyngeal carcinoma cases, 7,073 biopsy-verified benign hyperplasia cases, and 9,596 normal cases) collected across eight major nasopharyngeal carcinoma institutions (stage 1). Prospective external testing included 1,885 new images collected from 10 hospitals in regions with elevated disease rates (stage 2). A complete crossover multireader multicase experiment further included four specialists from top regional nasopharyngeal carcinoma facilities and 24 general otolaryngologists working in 24 varied primary care settings. This component analyzed 400 images to measure changes in diagnostic performance when clinicians used the STND system versus working independently in practical scenarios. Internal dataset images, obtained between January 1, 2017, and January 31, 2023, came from 15,521 people (9,033 men [58.2%] and 6,488 women [41.8%]; average age 47.6 years [IQR 38.4–56.8]). External validation used images from 945 individuals (538 men [56.9%] and 407 women [43.1%]; average age 45.2 years [IQR 35.2–55.2]). On internal data, STND reached an AUC of 0.99 (95% CI: 0.99–0.99) to distinguish normal nasopharynx from any abnormality (benign hyperplasia plus carcinoma) and an AUC of 0.99 (95% CI: 0.98–0.99) to identify malignant versus non-malignant cases. External validation yielded an AUC of 0.95 (95% CI: 0.94–0.96) for carcinoma detection, 91.6% sensitivity (95% CI: 89.3–93.5), and 86.1% specificity (95% CI: 84.1–87.9). In the reader experiment, AI support raised primary care otolaryngologists' overall accuracy by 7.9%, moving from 83.4% (95% CI: 80.1–86.7) unaided to 91.2% (95% CI: 88.6–93.9) with assistance ($P < 0.0001$). Per-image interpretation time also fell from 6.7 s (SD: 6.0) to 5.0 s (SD: 2.5) when using the model ($P = 0.034$). The developed deep learning platform shows strong promise for everyday clinical deployment in recognizing nasopharyngeal carcinoma via endoscopic imaging. It delivers clear advantages for frontline hospitals by improving specificity, reducing unnecessary biopsies, and preventing missed malignancies.

Keywords: Nasopharyngeal carcinoma, Artificial intelligence, Deep learning, Swin transformer, Endoscopic imaging, Computer-aided diagnosis

Access this article online

<https://smerpub.com/>

Received: 30 May 2024; Accepted: 27 June 2024

Copyright CC BY-NC-SA 4.0

How to cite this article: Eriksen K, Berg S, Dahl M. A National Multicenter Study on Artificial Intelligence-Based Detection of Nasopharyngeal Carcinoma from Endoscopic Images. *J Med Sci Interdiscip Res.* 2024;4(1):171-83. <https://doi.org/10.51847/8miluHMkSq>

Introduction

Nasopharyngeal carcinoma represents a malignancy of the mucosal lining in the nasopharynx. Over 70% of new diagnoses occur in Southeast Asia, making it a pressing concern in that region [1]. Successful therapy outcomes hinge on prompt recognition, yet most individuals seek medical attention only after the disease has progressed

because it is initially silent [2, 3]. Endoscopic nasal examination is the primary screening approach because of its simplicity, low cost, and rapid execution. Despite this, the close resemblance in local appearance between benign hyperplasia and carcinoma often hinders reliable differentiation, leading to delayed identification or missed diagnoses. Consequently, clinicians—especially those practicing in community hospitals with lower exposure—commonly order additional biopsies as a safeguard, thereby raising financial burdens, physical distress, and emotional strain on patients.

Research in context

Evidence before this study

PubMed was queried from its start through April 1, 2024, combining terms (“Nasopharyngeal Carcinoma” OR “Nasopharyngeal Cancer” OR “Benign Hyperplasia”) with (“Artificial Intelligence” OR “Deep Learning”) and (“Nasal Endoscopy” OR “Nasopharyngoscopy” OR “Endoscopy”). Four earlier reports addressed deep learning for carcinoma spotting and confirmed basic viability, albeit with inconsistent validation depth. Three operated within single institutions, while the remaining one studied only a modest patient group. Although AI tools have excelled in various retrospective analyses, routine clinical integration remains rare. Persistent obstacles, particularly around broad applicability, impede wider uptake. A truly useful system demands extensive multicentre training data, forward-looking external checks on endoscopic material, and thorough testing by practicing doctors in the exact environments where adoption is anticipated.

Added value of this study

Following an exhaustive survey of prior publications, this effort produced what appears to be the highest-performing and most transferable AI solution yet for separating nasopharyngeal carcinoma, benign hyperplasia, and normal mucosa under endoscopy, capitalizing on the broadest multicentre image collection documented to date. The resulting Swin Transformer-based Nasopharyngeal Diagnostic (STND) framework performed well when scrutinized through prospective external data. This work also marks the first effort to assess AI-supported endoscopic reading through a detailed multireader, multicase protocol, directly probing its usefulness in real clinical workflows. Results revealed notable gains in carcinoma identification, with particular relevance for under-resourced primary care contexts.

Implications of all the available evidence

The investigation delivers an exceptionally precise, stable, and widely adaptable AI platform to support the recognition of nasopharyngeal carcinoma from endoscopic images, validated with forward-collected samples from key epidemic zones nationwide. Clinician reader outcomes show that doctors, particularly those in basic healthcare facilities with limited background and infrastructure, can achieve clear improvements in accuracy and efficiency with STND, including faster case review.

Deep learning adoption has accelerated throughout healthcare, notably in head and neck oncology, promising meaningful progress in identification and care [4-6]. For nasopharyngeal carcinoma specifically, prior uses cover lesion separation from benign growths on MRI scans [7], tumor outlining in CT/MRI datasets [8-10], automatic malignancy grading via MRI [11], and interpretation of tissue slides [12]. Endoscopy-related deep learning explorations, however, remain infrequent [13, 14]. Significant examples include Li *et al.* [15] and the eNPM-DM architecture, built on 27,536 confirmed images from 7,951 patients, achieving 88.7% detection accuracy and a mean Dice score of 0.78 (SD: 0.24) [15]. Xu *et al.* [16] created an automated classifier using standard white-light and narrow-band modalities, achieving 94.9% accuracy and 0.99 AUC in cross-validation testing. Earlier team research presented a weakly semi-supervised Point SEGTR method that reduces the demand for fine-grained labeling in segmentation model training [17].

Previous efforts have demonstrated deep learning’s capacity to assist specialists in screening for nasopharyngeal carcinoma, yet several constraints persist. Foremost is the scarcity of cross-institutional evaluations. Endoscopy lacks uniform execution standards and quality oversight across sites, leading to wide differences in captured images stemming from operator proficiency, device configurations, and individual anatomical traits. Resulting inconsistencies in visual style, illumination, and perspectives challenge models trained in isolated settings when faced with new data. Furthermore, few large clinician-reader trials exist to verify true bedside impact. As medical AI moves toward operational use, validation within target practice environments becomes vital [18-20]. Such real-world appraisals confirm that tools meet genuine clinical requirements and genuinely improve performance. To

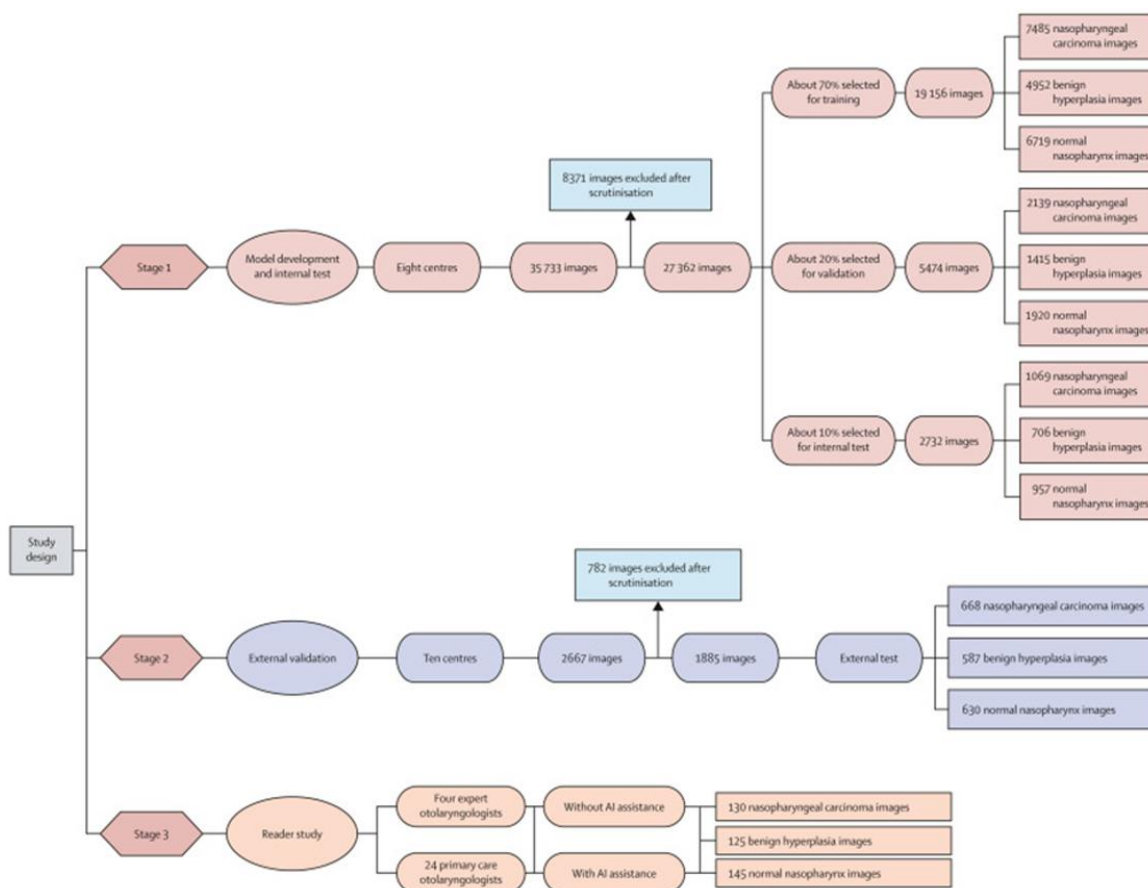
address these issues, the present research trained and evaluated a deep learning solution on diverse multicentre endoscopic datasets to classify carcinoma versus benign and normal states, while assessing its clinical utility through prospective validation and doctor-involved reader analysis.

Materials and Methods

Study design

This national, multicentre initiative to develop and validate a diagnostic model established the Swin Transformer-based Nasopharyngeal Diagnostic (STND) system for precise recognition of nasopharyngeal carcinoma, benign hyperplasia, and normal nasopharynx in everyday endoscopic practice. The comprehensive assessment of the system's robustness and clinical utility involved 42 medical centers across 24 provinces and

municipalities in China. Model construction and internal checks used adult patient (≥ 18 years) endoscopic images acquired at the Eye, Ear, Nose, and Throat Hospital of Fudan University and seven other busy regional nasopharyngeal carcinoma centers handling over 200 annual cases (stage 1). Subsequent evaluation occurred on prospectively assembled images from 10 separate hospitals in high-prevalence southeast zones that had not joined the development phase (stage 2). The concluding phase featured a fully crossed reader trial with four experienced otolaryngologists affiliated with prominent institutions specializing in nasopharyngeal carcinoma and 24 practicing general otolaryngologists from primary facilities to quantify the gains provided by STND (stage 3). All participating readers originated from institutions located along the eastern segment of the Hu Line [21] in China, an area that contains roughly 90% of the national population.



(a)

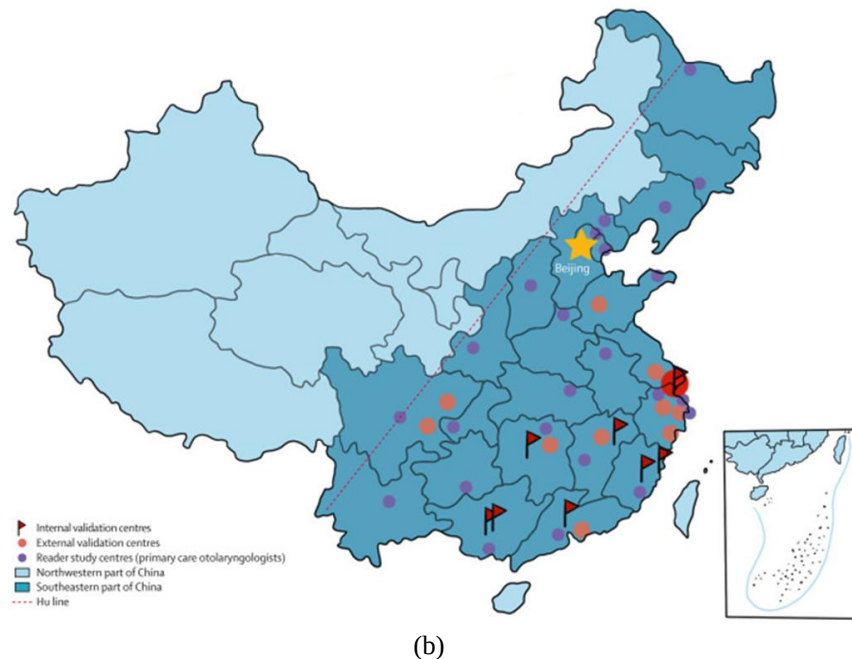


Figure 1. Study protocol and nationwide hospital locations.

The research protocol was approved by the Ethics Committee of the Eye and ENT Hospital of Fudan University (approval No. 2022052) and registered on ClinicalTrials.gov under identifier NCT05627310. All procedures followed the ethical standards of the Declaration of Helsinki because the retrospective internal dataset did not include any personally identifiable information; a waiver of informed consent was requested and granted by the institutional review boards of all eight participating centers. In contrast, every participant in the prospective external validation phase provided written informed consent. Financial support covered data-gathering and the procurement of computing equipment needed for model construction (**Figure 1**).

Model development and internal validation

To enable nasopharyngeal lesion classification, the STND system was constructed using 27,362 endoscopic images: 10,693 (39.1%) biopsy-confirmed nasopharyngeal carcinoma cases, 7,073 (25.8%) biopsy-confirmed benign hyperplasia cases, and 9,596 (35.1%)

normal nasopharynx images. We retrospectively assembled these data from eight medical centers across China. For internal validation, we randomly selected an additional subset of 2,732 images (1,069 nasopharyngeal carcinoma, 706 benign hyperplasia, and 957 normal) from the main dataset, maintaining roughly equal proportions (approximately 10%) from each contributing center.

The STND framework was built on an enhanced Swin Transformer backbone [22], augmented with a classification head comprising a global average pooling layer followed by a fully connected linear layer. The hidden layer channel dimension (C) in the initial stage was set to 128 (**Figure 1**). Input endoscopic images had height H and width W. Model training and inference were performed using the PyTorch framework (version 1.7.1) and Python 3.7.6. Optimization employed the Adam algorithm together with a cosine annealing learning rate schedule, starting at an initial rate of 0.0001 and a batch size of 16. Computations ran on a Tesla V100-SMX2 GPU (CUDA 10.1; Nvidia, Santa Clara, CA, USA).

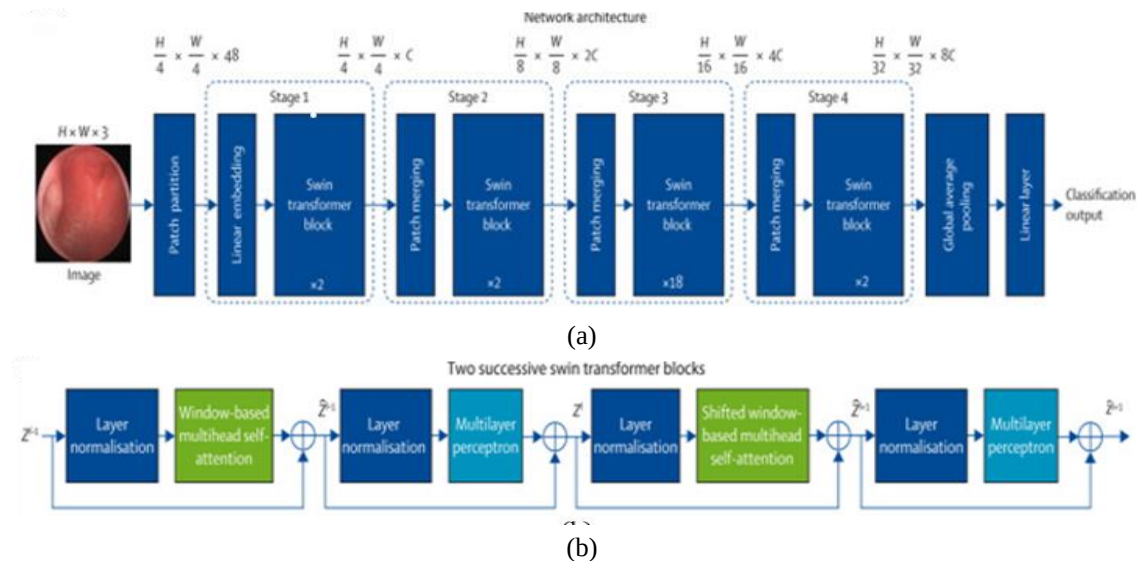


Figure 2. Overview of the STND framework for classifying nasopharyngeal conditions.

Prospective external validation

To test how well the STND system performs across different hospital settings and geographic areas in China, researchers conducted external testing on 1,885 nasopharyngeal endoscopic images. This set included 668 (35.4%) nasopharyngeal carcinoma images, 587 (31.1%) benign hyperplasia images, and 620 (32.9%) normal nasopharynx images. All images were prospectively collected from 10 independent hospitals that had not participated in the earlier model training phase.

Reader study with or without AI assistance

The investigation recruited 4 expert otolaryngologists and 24 primary care otolaryngologists for a fully crossed, multireader, multicase evaluation. The four specialists came from prominent nasopharyngeal carcinoma research institutions, had an average of 23 years of professional experience, held senior academic ranks, and demonstrated substantial expertise in clinical diagnosis. The 24 general otolaryngologists were selected from provinces along the eastern side of the Hu Line, with one participant per province, to mirror real-world application conditions for the STND system.

For the reader experiment, we randomly selected 400 endoscopic images from the 1,885-image external validation set, with selection proportions matching each center's overall contribution. Among these 400 images, there were 130 (32.5%) nasopharyngeal carcinoma cases, 125 (31.3%) benign hyperplasia cases, and 145 (37.3%) normal nasopharynx cases. Every reader evaluated and

classified all images twice—once without assistance and once with STND support—following a crossover design. We inserted a two-week washout interval between sessions to reduce memory effects [23]. No feedback on scores or correct answers was given after either session, preventing any influence from prior exposure. Before each reading round, all images were fully anonymized, preprocessed, and randomly ordered. Assessments were conducted via an electronic questionnaire containing 400 images. The system automatically logged the total time each reader spent across both sessions. All participants held active medical licenses in China and had at least 2 years of clinical practice. We also performed subgroup analyses among primary care otolaryngologists, stratified by years of experience, hospital level, and endoscopy equipment type, to examine AI's impact more closely.

Statistical analysis

We measured diagnostic performance relevant to everyday practice using a one-versus-rest strategy [24] to compute sensitivity, specificity, accuracy, and the area under the receiver operating characteristic curve (ROC-AUC). Metrics were derived separately for two binary classification tasks based on both model outputs and reader results: normal versus abnormal, and malignant versus non-malignant. In this setup, the normal-versus-abnormal task contrasted normal nasopharynx images against all abnormal images (nasopharyngeal carcinoma plus benign hyperplasia), while the malignant-versus-non-malignant task compared nasopharyngeal carcinoma

images against non-malignant ones (benign hyperplasia plus normal nasopharynx).

For standalone STND system evaluation, 95% confidence intervals for all metrics except ROC-AUC were calculated with Wilson's score method; the DeLong method was used for ROC-AUC. Sample size planning for the multireader multicase study determined that at least 257 cases would be required with 20 readers ($\alpha=0.025$, $\beta=0.20$; effect size 0.063), drawing on variance and effect estimates from a pilot analysis. Variance components included reader-treatment interaction variance (YTR) of 0.0019, case-treatment interaction variance (YTC) of 0.035, and residual error variance (YEps) of 0.069. We applied the Dorfman–Berbaum–Metz approach [25] to account for reader and case variability when comparing diagnostic performance between the two reading conditions. We conducted superiority testing for all reader study metrics by checking whether the lower bound of the two-sided 95% CI was greater than zero. We modeled readers and images as random effects, assuming a multivariate normal distribution. ROC-AUC values were estimated via the non-parametric trapezoidal rule. We analyzed differences in reading times with and without AI using the Wilcoxon signed-rank test after confirming a symmetric distribution of differences. All statistical computations were performed in R version 4.3.3 using the RJafroc package version 2.1.3 [26]. Superiority tests were one-sided, with statistical significance defined as $P < 0.025$ [27–29].

Role of the funding source

The study sponsors had no involvement in the design, data collection, data analysis, interpretation of results, or manuscript preparation.

Results and Discussion

From January 1, 2017, to January 31, 2023, we collected 35,733 endoscopic images from eight hospitals for the internal phase. After applying the study's inclusion and exclusion standards, 8,371 images were removed, leaving 27,362 images from 15,521 patients (9,033 men [58.2%] and 6,488 women [41.8%]; mean age 47.6 years [IQR 38.4–56.8]). During the prospective external validation period (October 1 to December 31, 2023), we collected 2,667 images over three months, of which 782 were excluded after screening. This resulted in 1,885 images from 945 patients (538 men [56.9%] and 407 women [43.1%]; mean age 45.2 years [IQR 35.2–55.2]). All images were captured at 96 dpi resolution with dimensions exceeding 1024×768 pixels. Benign hyperplasia cases were defined according to biopsy findings as non-specific chronic inflammation, excluding benign tumors.

In the internal validation set, the system achieved an AUC of 0.99 (95% CI: 0.99–0.99) for distinguishing normal versus abnormal images and an AUC of 0.99 (95% CI: 0.98–0.99) for malignant versus non-malignant classification (**Table 1**). Sensitivity for normal-versus-abnormal separation was 93.1% (95% CI: 91.9–94.2) with specificity of 98.6% (95% CI: 97.7–99.2). For malignant-versus-non-malignant separation, sensitivity was 95.0% (95% CI: 93.5–96.1), and specificity was 94.7% (95% CI: 93.5–95.7).

Table 1. Sensitivity, specificity, and AUC of the proposed STND in the internal and external validation

	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	ROC-AUC (95% CI)
Internal dataset				
Normal vs abnormal	93.1 (91.9–94.2)	98.6 (97.7–99.2)	95.1 (94.2–95.8)	0.99 (0.99–0.99)
Malignant vs non-malignant	95.0 (93.5–96.1)	94.7 (93.5–95.7)	94.8 (93.9–95.6)	0.99 (0.98–0.99)
External dataset				
Normal vs abnormal	91.6 (89.9–93.0)	88.6 (85.8–90.8)	90.6 (89.2–91.8)	0.95 (0.94–0.96)
Malignant vs non-malignant	91.6 (89.3–93.5)	86.1 (84.1–87.9)	88.1 (86.5–89.5)	0.95 (0.94–0.96)

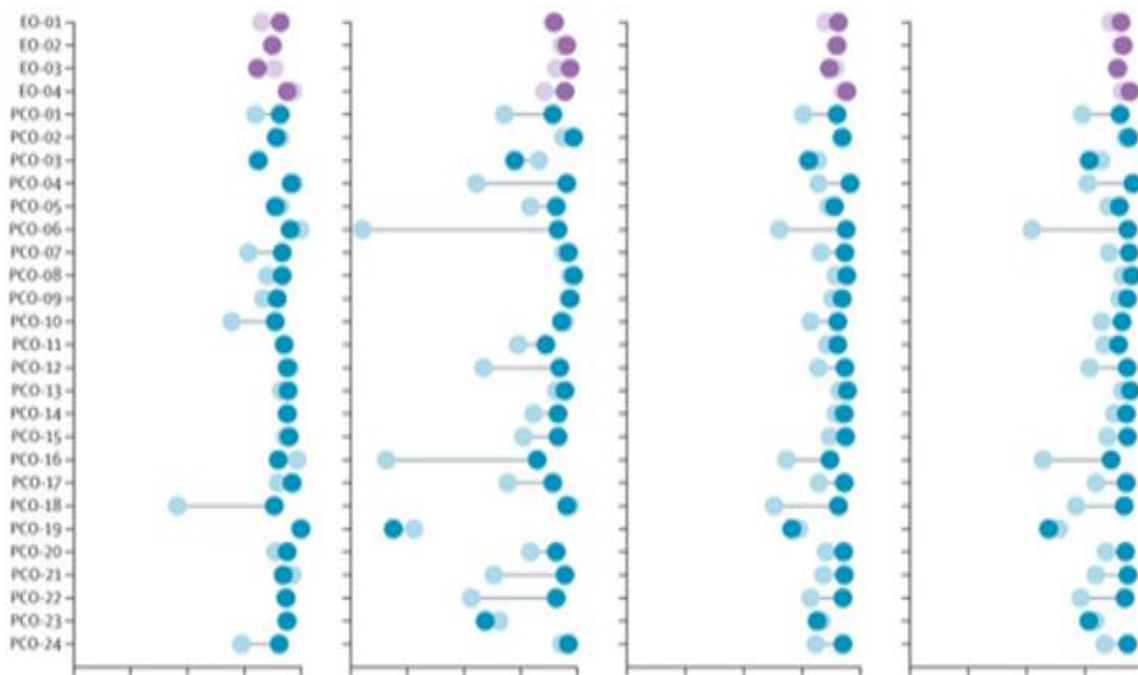
Across stages 1 and 2, the system underwent internal assessment using data sourced from 8 hospitals and external assessment using data from 10 hospitals. The model demonstrated excellent accuracy when tested on the internal dataset. However, performance modestly

declined when evaluated on the external dataset. While the system maintained comparatively strong sensitivity in external tests, specificity, along with several other performance indicators, declined when confronted with previously unseen data. We applied Wilson's score

method to compute 95% confidence intervals for sensitivity, specificity, and accuracy. We used the DeLong method to calculate the ROC-AUC. AUC: area under the curve. ROC-AUC: area under the conventional receiver operating characteristic curve. STND: Swin Transformer-based Nasopharyngeal Diagnostic.

In the external validation cohort, the model attained AUC values of 0.95 (95% CI: 0.94–0.96) for distinguishing normal versus abnormal images and 0.952 (95% CI: 0.94–0.96) for separating malignant versus non-malignant images (**Table 1**). When examined individually across 10 external centers, AUCs ranged from 0.90 to 0.98 for the normal-versus-abnormal task and from 0.86 to 0.97 for the malignant-versus-non-malignant task.

When evaluating normal versus abnormal cases, expert otolaryngologists showed only modest improvements in most performance measures (**Table 2; Figure 3**). Specificity rose by 4.1% (95% CI: –0.9 to 9.1; $P = 0.043$), accuracy by 1.2% (95% CI: –3.4 to 5.8; $P = 0.25$), and AUC by 0.02 (95% CI: –0.01 to 0.05; $P = 0.084$). Sensitivity, however, showed a minor reduction of 0.4% (95% CI: –9.7 to 8.9; $P = 0.45$). By comparison, primary care otolaryngologists experienced substantially larger benefits. Their sensitivity improved to 92.9% (95% CI: 90.4–95.4), specificity to 88.6% (95% CI: 81.9–95.4), accuracy to 91.2% (95% CI: 88.6–93.9), and AUC to 0.91 (95% CI: 0.87–0.94).



(a)

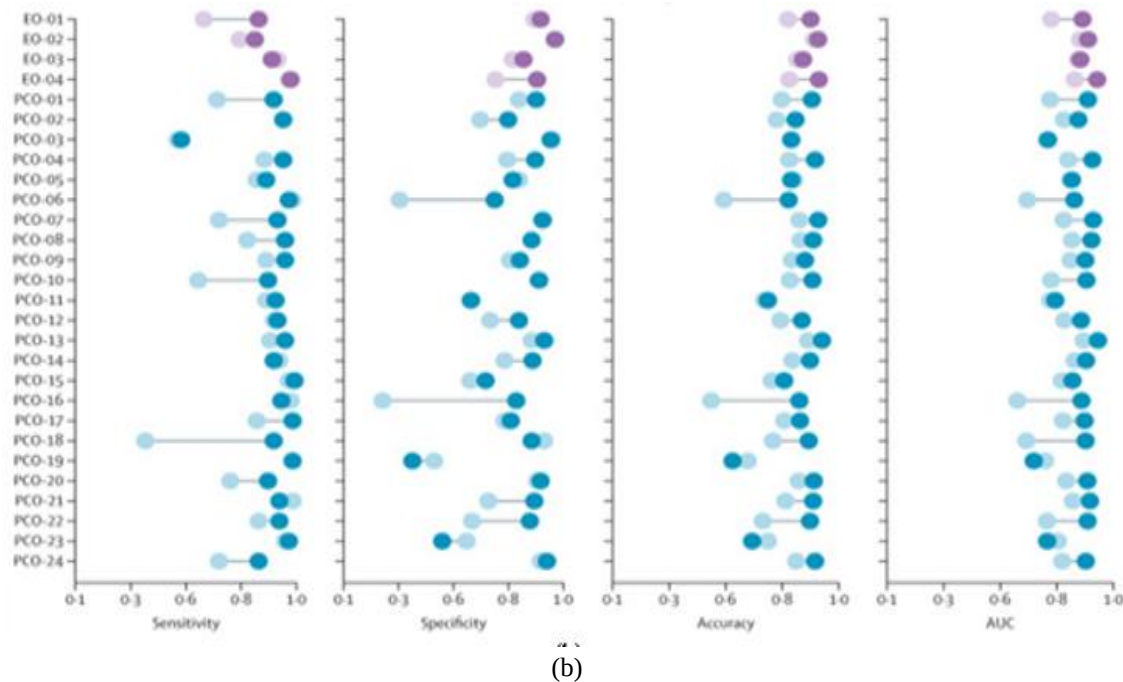


Figure 3. Performance change for each reader with AI assistance

Table 2. Sensitivity, specificity, and AUC of expert otolaryngologists and primary care otolaryngologists with and without AI assistance in the reader study.

Classification Task	Evaluation Group	Sensitivity (95% CI)	Sensitivity Change	P-value	Specificity (95% CI)	Specificity Change	P-value	Overall Accuracy (95% CI)	Accuracy Change	P-value	ROC-AUC (95% CI)	ROC-AUC Change	P-value
Normal vs. Abnormal	AI Model Only	92.9 (89.1–95.5)	–	–	90.3 (84.4–94.2)	–	–	92.0 (88.9–94.3)	–	–	0.92 (0.89–0.95)	–	–
	Expert Otolaryngologists		–0.4 (–9.7 to 8.9)	0.45		4.1 (–0.9 to 9.1)	0.043		1.2 (–3.4 to 5.8)	0.25		0.02 (–0.01 to 0.05)	0.084
	Without AI	89.8 (82.1–97.5)	–	–	90.7 (86.4–95.0)	–	–	90.1 (86.1–94.2)	–	–	0.90 (0.87–0.93)	–	–
	Assisted by AI	89.4 (82.0–96.8)	–	–	94.8 (90.9–98.7)	–	–	91.3 (87.3–95.3)	–	–	0.92 (0.89–0.95)	–	–
Primary Care Otolaryngologists			4.2 (0.0 to 8.3)	0.024		14.5 (5.4 to 23.6)	0.0016		7.9 (4.6 to 11.2)	<0.0001		0.09 (0.05 to 0.14)	< 0.0001
	Without AI	88.7 (83.8–93.7)	–	–	74.2 (64.3–84.1)	–	–	83.4 (80.1–86.7)	–	–	0.81 (0.77–0.86)	–	–

		92.9 (90.4– 95.4)	–	–	88.6 (81.9– 95.4)	–	–	91.2 (88.6– 93.9)	–	–	0.91 (0.87– 0.94)	–	–
Malignant vs. Non- malignant	AI Model Only	93.1 (87.4– 96.3)	–	–	88.9 (84.6– 92.1)	–	–	90.3 (86.9– 92.8)	–	–	0.91 (0.88– 0.94)	–	–
	Expert Otolaryngologists		6.7 (–10.6 to 24.0)	0.15		6.1 (–5.7 to 17.9)	0.10		6.2 (–1.7 to 14.0)	0.044		0.06 (–0.02 to 0.15)	0.044
	Without AI	82.1 (57.5– 100.0)	–	–	84.0 (67.6– 100.0)	–	–	83.4 (76.2– 90.5)	–	–	0.83 (0.75– 0.91)	–	–
	Assisted by AI	88.9 (79.1– 98.6)	–	–	90.1 (82.1– 98.1)	–	–	89.6 (85.3– 93.9)	–	–	0.90 (0.85– 0.94)	–	–
	Primary Care Otolaryngologists		9.2 (3.5 to 14.9)	0.0013		7.7 (1.4 to 14.0)	0.0096		8.1 (4.2 to 12.1)	0.0001		0.08 (0.05 to 0.12)	<0.0001
	Without AI	82.3 (74.9– 89.7)	–	–	73.0 (64.8– 81.3)	–	–	76.0 (71.5– 80.4)	–	–	0.78 (0.74– 0.81)	–	–
	Assisted by AI	91.5 (87.0– 96.0)	–	–	80.7 (74.2– 87.2)	–	–	84.1 (79.9– 88.2)	–	–	0.86 (0.83– 0.89)	–	–

A fully crossed multireader multicase evaluation formed stage 3 of the study. It involved four expert otolaryngologists from specialized nasopharyngeal carcinoma centers and 24 primary care otolaryngologists from diverse primary hospitals. The purpose was to determine how the Swin Transformer-based Nasopharyngeal Diagnostic system influenced clinicians' diagnostic performance. We analyzed differences in diagnostic accuracy using the Dorfman–Berbaum–Metz method. AI = artificial intelligence. AUC = area under the curve.

When classifying malignant versus non-malignant lesions, both expert and primary care groups achieved notable gains in accuracy and AUC (**Table 2; Figure 3**). Experts improved accuracy by 6.2% (95% CI: –1.7 to 14.0; $P = 0.044$) and AUC by 0.06 (95% CI: –0.02 to 0.15; $P = 0.044$). Primary care otolaryngologists benefited even more, with accuracy rising by 8.1% (95% CI: 4.2 to 12.1; $P = 0.0001$) and AUC by 0.08 (95% CI: 0.05 to 0.12; $P < 0.0001$).

Subgroup analysis among primary care otolaryngologists with varying years of nasal endoscopy experience revealed clear benefits from AI support. Those with fewer than 10 years of experience achieved major gains

in the normal-versus-abnormal task: specificity increased from 72.5% (95% CI: 59.7–85.3) to 90.2% (95% CI: 84.0–96.4), a rise of 17.7% (95% CI: 5.6–29.8; $P = 0.0039$); accuracy rose from 84.7% (95% CI: 80.8–88.7) to 92.2% (95% CI: 89.6–94.9), an increase of 7.5% (95% CI: 3.9–11.1; $P = 0.0002$); and AUC improved from 0.82 (95% CI: 0.77–0.88) to 0.918 (95% CI: 0.89–0.95), a gain of 0.097 (95% CI: 0.05–0.15; $P = 0.0007$). Clinicians with 10 or more years of experience showed meaningful improvement in sensitivity, rising from 85.7% (95% CI: 76.8–94.6) to 92.3% (95% CI: 88.9–95.6). Comparable patterns of benefit appeared in the malignant-versus-non-malignant task.

The research team measured the time needed for STND to analyze 400 images on a workstation equipped with an Intel Xeon Platinum 8255C @ 2.50GHz CPU and a Tesla V100-SMX2 GPU. Total processing time was 10.1 seconds, yielding an average of 0.025 seconds per image. This speed supports real-time predictions during nasopharyngoscopy examinations. In addition, readers' interpretation time per image decreased significantly with AI assistance (mean 5.0 s [SD: 2.5]; median 4.0 s [IQR: 3.6–5.3]) compared with unaided reading (mean 6.7 s [SD: 6.0]; median 5.0 s [IQR: 4.0–6.7]; $P = 0.034$).

The exact causes of nasopharyngeal carcinoma are not fully understood, making prompt identification essential for successful treatment and improved patient prognosis. Nasal endoscopy is the primary, most practical screening method currently available. In routine practice, otolaryngologists follow a two-stage process: first, identifying any suspicious areas in the nasopharynx; then, determining whether they represent benign hyperplasia or cancer, often requiring biopsy for confirmation. This workflow depends on two key distinctions—normal versus abnormal, and malignant versus non-malignant. In primary care settings, less experienced clinicians struggle more with the second step than the first. They frequently adopt a cautious approach, labeling normal cases as abnormal to avoid missing malignancies. This results in high sensitivity but reduced specificity. To tackle these issues, the current study created a detection algorithm trained on more than 27,362 images collected from eight major nasopharyngeal carcinoma centers. The resulting model achieved superior classification accuracy, sensitivity, and specificity compared with earlier reports [13, 15, 16]. Given the substantial variability in endoscopic images caused by differences in clinician technique, equipment settings, and patient anatomy (affecting image style, brightness, and angles), the team conducted external validation across ten hospitals in high-incidence areas to test robustness on diverse real-world data. This large-scale multicentre external validation represents, to our knowledge, the first of its kind for AI-supported nasopharyngeal carcinoma screening.

To realistically evaluate performance in resource-limited primary hospitals and mimic actual clinical use, the study included the largest reader investigation conducted to date. It recruited four highly experienced expert otolaryngologists from leading regional centers and 24 primary care otolaryngologists. Application of the STND system produced meaningful improvements in diagnostic performance for both groups, though the magnitude of these improvements varied. Experts saw modest benefits in the normal-versus-abnormal task, including a 1.2% rise in accuracy, and more noticeable gains in the malignant-versus-non-malignant task, where accuracy increased by 6.2% ($P < 0.05$). Primary care clinicians recorded large improvements across all key metrics—sensitivity, specificity, accuracy, and AUC—for both abnormality detection and malignancy differentiation. Specificity rose by 14.5% ($P < 0.01$), and AUC increased from 0.81 to 0.91 ($P < 0.0001$) when detecting

abnormalities. With AI support, primary care otolaryngologists reached specificity levels close to those of experts (88.6% versus 90.7% for normal-versus-abnormal; 80.7% versus 84.0% for malignant-versus-non-malignant). These gains can facilitate more precise differential diagnosis and help decrease unnecessary biopsies in daily practice.

Nevertheless, individual responses among primary care doctors varied. For example, primary care otolaryngologists PCO-06 and PCO-16 showed large increases in specificity, consistent with less experienced clinicians initially over-recommending biopsies for safety; AI assistance helped them reduce such procedures—one of the system's main aims. Conversely, PCO-19 showed decreases in specificity, accuracy, and AUC, possibly due to lower confidence in the AI's recommendations.

We further subgroup-analyzed primary care otolaryngologists by years of experience, hospital grade, and endoscopy type. Participants were divided into junior (< 10 years) and senior (≥ 10 years) groups. Without assistance, junior clinicians tended toward more aggressive decision-making, achieving higher sensitivity (91.7% versus 85.7% for normal-versus-abnormal; 87.8% versus 76.8% for malignant-versus-non-malignant) but lower specificity (72.5% versus 75.9% for normal-versus-abnormal; 69.5% versus 76.6% for malignant-versus-non-malignant). With STND support, juniors outperformed seniors across all measures, especially specificity, with increases of 17.7% for normal-versus-abnormal and 11.1% for malignant-versus-non-malignant. When stratified by hospital grade, primary care doctors in grade II hospitals initially performed better unaided, while those in grade III hospitals showed larger gains with AI. This pattern aligned with differences in experience (mean 11.5 years [SD: 4.52] in grade III versus 13.6 years [SD: 8.05] in grade II), suggesting that clinical experience matters more than hospital level. Overall, junior otolaryngologists appeared more open to AI assistance, whereas seniors relied more on personal expertise, potentially limiting additional benefit from the tool. Regarding equipment, flexible endoscopes yielded higher AUCs for abnormality detection, while rigid endoscopes performed better for malignancy identification. Taken together, widespread deployment of STND in primary hospitals could substantially raise the standard of nasopharyngeal carcinoma screening, cut

unnecessary invasive tests, and support earlier, more appropriate interventions.

For responsible clinical integration of STND, the study emphasizes proactive measures to protect patient privacy and minimize algorithmic bias, thereby fostering confidence among clinicians and patients. The study established comprehensive ethical frameworks and protocols to safeguard data confidentiality. All training and operational data were de-identified and encrypted. We implemented access controls, continuous monitoring, and secure storage solutions that meet or exceed current standards. Hospitals are advised to run the system in offline mode after deployment to eliminate external network risks. To counter potential biases arising from training data, dataset curation deliberately incorporated diversity in hospital settings, patient demographics, and equipment types. Large, balanced datasets help prevent unfair performance across groups. Post-deployment, ongoing performance surveillance will detect and correct any emerging biases, with model updates released as needed. Hospitals adopting STND are encouraged to form local ethics committees, including clinicians, data scientists, and ethicists, to oversee issues of patient autonomy, consent, and transparency and ensure continued ethical compliance.

Several limitations should be noted. Although the model achieved 95.0% sensitivity internally and 91.6% externally for detecting nasopharyngeal carcinoma, sensitivity varied across centers. This is probably attributable to variations in endoscopic equipment, imaging protocols, and overall image quality. Centers employed different endoscope models with varying resolution and contrast settings; even identical equipment can produce inconsistent results due to adjustments in exposure or filtering. Noise, artifacts, and maintenance issues can further obscure important visual details. As the national quality control center for otorhinolaryngology, the team intends to disseminate best practices to primary hospitals nationwide to standardize procedures and elevate image quality, thereby supporting improved national diagnostic standards. Second, STND relies on visible surface features in static endoscopic images [30]. Very early submucosal nasopharyngeal carcinomas without obvious mucosal changes would therefore evade detection. Emerging modalities such as Raman spectroscopy [31], photoacoustic imaging [32], and immunofluorescence [33] can probe beneath the surface but remain largely experimental. Future combination of these technologies with STND could enable more

comprehensive multimodality screening. Third, the current system analyses static images, which contain less information than dynamic video recordings. Although video-based models have been explored previously, they have been restricted to single-center data because most facilities do not routinely store videos. Prospective multicentre studies using video data are planned to advance more sophisticated algorithms.

Conclusion

In summary, STND offers effective assistance to otolaryngologists during nasopharyngeal carcinoma screening, with particular value in primary care settings. It excels at differentiating benign hyperplasia from malignancy, thereby reducing unwarranted invasive procedures. By limiting unnecessary interventions, the system may decrease patient waiting times, reduce healthcare expenditures, improve outcomes, and ease workload on medical teams. For safe deployment, offline operation is recommended to protect data security and comply with privacy regulations. Continuous oversight by ethics committees will support transparent, equitable use. The creation and implementation of STND supports the WHO's objectives in the 2030 Agenda for Sustainable Development, which stresses the need for tools enabling early cancer detection and screening. Broad adoption in primary hospitals across China and, potentially, throughout Southeast Asia could meaningfully expand access to timely treatment, improve survival rates, and contribute to global efforts against nasopharyngeal carcinoma.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Chen YP, Chan ATC, Le QT, Blanchard P, Sun Y, Ma J, et al. Nasopharyngeal carcinoma. *Lancet*. 2019;394:64-80.
2. Au KH, Ngan RKC, Ng AWY, Poon DMC, Ng WT, Yuen KT, et al. Treatment outcomes of nasopharyngeal carcinoma in modern era after

- intensity modulated radiotherapy (IMRT) in Hong Kong: a report of 3328 patients (HKNPCSG 1301 study). *Oral Oncol.* 2018;77:16-21.
3. Liang H, Xiang YQ, Lv X, Guo SS, Yang J, Liu YW, et al. Survival impact of waiting time for radical radiotherapy in nasopharyngeal carcinoma: a large institution-based cohort study from an endemic area. *Eur J Cancer.* 2017;73:48-60.
 4. Abdel Razek AAK, Khaled R, Helmy E, Nada N, Elsorogy L, El-Said M. Artificial intelligence and deep learning of head and neck cancer. *Magn Reson Imaging Clin N Am.* 2022;30:81-94.
 5. Rokhshad R, Salehi SN, Yavari A, Ziaei H, Sadeghi M, Mohammadzadeh M, et al. Deep learning for diagnosis of head and neck cancers through radiographic data: a systematic review and meta-analysis. *Oral Radiol.* 2024;40:1-20.
 6. Wang Z, Fang M, Zhang J, Xu M, Dong D, Tian J, et al. Radiomics and deep learning in nasopharyngeal carcinoma: a review. *IEEE Rev Biomed Eng.* 2024;17:118-35.
 7. Wong LM, King AD, Ai QYH, Mo FKF, Huang Y, Vlantis AC, et al. Convolutional neural network for discriminating nasopharyngeal carcinoma and benign hyperplasia on MRI. *Eur Radiol.* 2021;31:3856-63.
 8. Daoud B, Morooka K, Kurazume R, Le TN, Maki T, Kobayashi K, et al. 3D segmentation of nasopharyngeal carcinoma from CT images using cascade deep learning. *Comput Med Imaging Graph.* 2019;77:101644.
 9. Lin L, Dou Q, Jin YM, Zhou GQ, Tang YQ, Chen WL, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. *Radiology.* 2019;291:677-86.
 10. Zhao L, Lu Z, Jiang J, Zhou Y, Li X, Li J, et al. Automatic nasopharyngeal carcinoma segmentation using fully convolutional networks with auxiliary paths on dual-modality PET-CT images. *J Digit Imaging.* 2019;32:462-70.
 11. Yang Q, Guo Y, Ou X, Xie Y, Gong G, Yin Y, et al. Automatic T staging using weakly supervised deep learning for nasopharyngeal carcinoma on MR images. *J Magn Reson Imaging.* 2020;52:1074-82.
 12. Diao S, Hou J, Yu H, Zhao Y, Zhang H, Zhang W, et al. Computer-aided pathologic diagnosis of nasopharyngeal carcinoma based on deep learning. *Am J Pathol.* 2020;190:1691-700.
 13. Wang SX, Li Y, Zhu JQ, Chen XM, Liu ZY, Zhang L, et al. The detection of nasopharyngeal carcinomas using a neural network based on nasopharyngoscopic images. *Laryngoscope.* 2024;134:127-35.
 14. He Z, Zhang K, Zhao N, Li Y, Wang J, Liu S, et al. Deep learning for real-time detection of nasopharyngeal carcinoma during nasopharyngeal endoscopy. *iScience.* 2023;26:107463.
 15. Li C, Jing B, Ke L, Li Y, Wang X, Chen Y, et al. Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies. *Cancer Commun (Lond).* 2018;38:59.
 16. Xu J, Wang J, Bian X, Li Y, Zhang H, Liu W, et al. Deep learning for nasopharyngeal carcinoma identification using both white light and narrow-band imaging endoscopy. *Laryngoscope.* 2022;132:999-1007.
 17. Shi Y, Wang H, Ji H, Li Q, Zhang R, Wu J, et al. A deep weakly semi-supervised framework for endoscopic lesion segmentation. *Med Image Anal.* 2023;90:102973.
 18. Tang R, Liang H, Guo Y, Liu Z, Wang Y, Chen X, et al. Pan-mediastinal neoplasm diagnosis via nationwide federated learning: a multicentre cohort study. *Lancet Digit Health.* 2023;5:e560-70.
 19. Maas MHJ, Neumann H, Shirin H, Gralnek I, Bisschops R, van der Laan J, et al. A computer-aided polyp detection system in screening and surveillance colonoscopy: an international, multicentre, randomised, tandem trial. *Lancet Digit Health.* 2024;6:e157-65.
 20. Jiang X, Hoffmeister M, Brenner H, Jansen L, Bläker H, Chang-Claude J, et al. End-to-end prognostication in colorectal cancer by deep learning: a retrospective, multicentre study. *Lancet Digit Health.* 2024;6:e33-43.
 21. Qi W, Liu S, Zhao M, Wang H, Li X, Zhang Y, et al. China's different spatial patterns of population growth based on the "Hu Line". *J Geogr Sci.* 2016;26:1611-25.
 22. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2021.
 23. Dong Z, Wang J, Li Y, Zhang H, Liu Y, Chen X, et al. Explainable artificial intelligence incorporated

- with domain knowledge diagnosing early gastric neoplasms under white light endoscopy. *NPJ Digit Med.* 2023;6:64.
24. Milea D, Najjar RP, Zhubo J, Ting DSW, Vasseneix C, Wong TY, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med.* 2020;382:1687-95.
 25. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Invest Radiol.* 1992;27:723-31.
 26. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. *Med Phys.* 2004;31:2313-30.
 27. Mansournia MA, Nazemipour M, Etminan M. P-value, compatibility, and S-value. *Glob Epidemiol.* 2022;4:100085.
 28. Mansournia MA, Nazemipour M. Recommendations for accurate reporting in medical research statistics. *Lancet.* 2024;403:611-2.
 29. Greenland S, Mansournia MA, Joffe M. To curb research misreporting, replace significance and confidence by compatibility: a preventive medicine golden jubilee article. *Prev Med.* 2022;164:107127.
 30. Chan KCA, Woo JKS, King A, Zee BCY, Lam WKJ, Chan SL, et al. Analysis of plasma Epstein–Barr virus DNA to screen for nasopharyngeal cancer. *N Engl J Med.* 2017;377:513-22.
 31. Shu C, Yan H, Zheng W, Wang Y, Li X, Liu Z, et al. Deep learning-guided fiberoptic Raman spectroscopy enables real-time in vivo diagnosis and assessment of nasopharyngeal carcinoma and post-treatment efficacy during endoscopy. *Anal Chem.* 2021;93:10898-906.
 32. Lin Y, Qiu T, Lan Y, Yang H, Li J, Chen G, et al. Multi-modal optical imaging and combined phototherapy of nasopharyngeal carcinoma based on a nanoplatform. *Int J Nanomedicine.* 2022;17:2435-46.
 33. Zhu J, Kamara S, Cen D, Tang H, Chen Y, Li X, et al. Generation of novel affibody molecules targeting the EBV LMP2A N-terminal domain with inhibiting effects on the proliferation of nasopharyngeal carcinoma cells. *Cell Death Dis.* 2020;11:213.