Cancer Screening and Prevention **2026** | Epub
DOI: 10.14218/CSP.2026.00010

Original Article

Artificial Intelligence-assisted Endoscopic Diagnosis of Upper Gastrointestinal Precancerous Lesions: A Systematic Review and Meta-analysis

Yang Bai[1], Danqi Huang[2], Jingyi Liu[2], Yibei Li[2], Jingbo Zhai[3], Tian Gan[4*] and Jiang Li[2*]

[1]*The Second Hospital & Clinical Medical School, Lanzhou University, Lanzhou, Gansu, China;* [2]*Office of Cancer Screening, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China;* [3]*School of Public Health, Tianjin University of Traditional Chinese Medicine, Tianjin, China;* [4]*School of Computer Technology and Sciences, Shandong University, Qingdao, Shandong, China*

Received: June 10, 2026 | **Revised:** July 18, 2026 | **Accepted:** August 21, 2026 | **Published online:** September 10, 2026

Abstract

Background and objectives: Upper gastrointestinal precancerous lesions are important targets for cancer screening and prevention. This study evaluated the diagnostic accuracy of artificial intelligence (AI)-assisted endoscopy for upper gastrointestinal precancerous lesions.

Methods: PubMed, Embase, MEDLINE, Cochrane Library, and IEEE Xplore were searched from inception to March 31, 2026. Studies reporting sensitivity and specificity, or sufficient data for their calculation, were included. Two reviewers independently extracted data. Pooled sensitivity, specificity, and the area under the receiver operating characteristic curve were estimated using a bivariate random-effects model.

Results: Fifteen studies comprising 24 datasets and 19,578 records were included. Pooled sensitivity and specificity were 0.930 (95% confidence interval (CI): 0.886–0.958) and 0.944 (95% CI: 0.906–0.967), respectively; the diagnostic odds ratio was 223.79 (95% CI: 92.24–542.94), and the area under the receiver operating characteristic curve was 0.9661. Substantial heterogeneity was observed (bivariate I^2 = 73%). Among the variables formally evaluated by meta-regression, imaging modality was associated with heterogeneity (global test P = 0.005); image-enhanced endoscopy had higher sensitivity than white-light endoscopy (0.960 vs. 0.875). No significant publication bias was detected by Egger's test (P = 0.833), Begg's test (P = 0.641), or Deeks' funnel plot asymmetry test (P = 0.053).

Conclusions: AI-assisted endoscopy shows high diagnostic accuracy for upper gastrointestinal precancerous lesions. Multicenter prospective studies are needed for validation.

Introduction

Upper gastrointestinal (UGI) cancers, principally gastric and esophageal cancers, collectively impose a substantial global health burden.[1,2] Despite overall improvements in treatment modalities, the 5-year survival rate for advanced-stage gastric and esophageal cancers remains below 30%, largely because the majority of patients are diagnosed at late stages when curative resection is no longer feasible.[3] In contrast, when detected at an early stage, the 5-year survival rate can exceed 90%, underscoring the importance of early detection strategies.[4]

Both gastric and esophageal cancers develop through well-defined precancerous cascades that offer critical windows for intervention. Gastric carcinogenesis follows the Correa cascade: chronic *Helicobacter pylori* gastritis progresses to chronic atrophic gastritis (CAG), gastric intestinal metaplasia (GIM), dysplasia, and finally adenocarcinoma. Patients with CAG or GIM carry an

Keywords: Artificial intelligence; Deep learning; Upper gastrointestinal tract; Precancerous lesions; Endoscopy; Meta-analysis; Cancer screening and prevention.
***Correspondence to:** Jiang Li, Office of Cancer Screening, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China. ORCID: https://orcid.org/0000-0001-8244-0349. Tel: +86-13261635911, E-mail: lij@cicams.ac.cn; Tian Gan, School of Computer Technology and Sciences, Shandong University, Qingdao, Shandong 266237, China. ORCID: https://orcid.org/0000-0003-3197-5698. Tel: +86-15564850500, E-mail: gantian@sdu.edu.cn
How to cite this article: Bai Y, Huang D, Liu J, Li Y, Zhai J, Gan T, *et al*. Artificial Intelligence-assisted Endoscopic Diagnosis of Upper Gastrointestinal Precancerous Lesions: A Systematic Review and Meta-analysis. *Cancer Screen Prev*. Published online: Sep 10, 2026. DOI: https://doi.org/10.14218/CSP.2026.00010.

annual gastric cancer risk of 0.1–0.25%, rising to > 6% for high-grade dysplasia.[5,6] In the esophagus, Barrett's esophagus, a form of intestinal metaplasia driven by chronic reflux, confers a 30–60-fold increased risk of esophageal adenocarcinoma.[7] Esophageal squamous intraepithelial neoplasia (ESIN) represents the precursor spectrum for esophageal squamous cell carcinoma, with annual progression rates of approximately 1% for low- to moderate-grade dysplasia.[8] Despite the potential for curative intervention, endoscopic recognition of these lesions remains suboptimal.

Esophagogastroduodenoscopy with biopsy is the standard approach to surveillance, yet conventional white-light endoscopy (WLE) has significant limitations. Detection of CAG and GIM relies on subtle mucosal changes (pallor, loss of rugal folds, irregular surface pattern) that are easily overlooked, with reported sensitivities of only 50–70% and substantial interobserver variability.[9] The Prague C & M criteria provide a standardized endoscopic grading system for Barrett's esophagus,[10] but dysplastic lesions may still be subtle against a heterogeneous background. ESIN is even more challenging to detect because it frequently lacks visible abnormalities and requires Lugol's chromoendoscopy, which is time-consuming, poorly tolerated, and limited in availability.[11] Consequently, a substantial proportion of precancerous lesions are missed during routine esophagogastroduodenoscopy, which may contribute to delayed diagnosis and preventable cancer progression.

Artificial intelligence (AI), particularly deep learning using convolutional neural networks (CNNs) and vision transformers, can support real-time identification of subtle mucosal abnormalities and may improve consistency relative to unaided human interpretation.[12] Retrospective studies report areas under the receiver operating characteristic curves (AUCs) > 0.90 for early gastric cancer, GIM, CAG, and Barrett's neoplasia.[13] Despite advances from CNNs to transformer architectures and real-time video integration,[14,15] the comparative performance and generalizability of AI across UGI precancerous lesions remain incompletely characterized.

Prior meta-analyses have evaluated AI for individual lesion types (e.g., early gastric cancer or Barrett's neoplasia), but evidence comparing diagnostic performance across different lesion categories (ESIN, Barrett's dysplasia, CAG, GIM, and gastric dysplasia), AI model architectures (CNN versus transformer), and imaging modalities (WLE versus image-enhanced endoscopy) remains limited. To address these gaps, we conducted a systematic review and meta-analysis to evaluate pooled diagnostic accuracy, perform subgroup analyses by lesion type, architecture, and modality, and assess heterogeneity and publication bias. Our findings provide an evidence base to guide clinical integration of AI-assisted endoscopic surveillance.

Given that precancerous lesions are actionable targets of organized cancer-screening programs, such a synthesis is directly relevant to setting risk-stratified screening intervals and allocating resources for cancer prevention.

Materials and methods

This systematic review and meta-analysis was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (the completed PRISMA checklist is provided in Supplementary File 1).[16] The study protocol was prospectively registered with PROSPERO (registration number: CRD420261349387).

Search strategy

Two authors independently searched electronic databases (PubMed, Ovid MEDLINE(R) ALL, Embase, Cochrane Library, and IEEE Xplore) from inception to March 31, 2026. We used the following keywords and corresponding Medical Subject Headings (MeSH) terms: "upper gastrointestinal tract", "precancerous conditions", "Barrett's esophagus", "intestinal metaplasia", "atrophic gastritis", "dysplasia", "artificial intelligence", "deep learning", "convolutional neural network", "computer-aided diagnosis", "endoscopy", "gastroscopy", "sensitivity", "specificity", and "accuracy". The full search strategy is shown in Supplementary File 2. The search was limited to English-language publications, and the references of the included studies and related reviews were manually searched to identify additional eligible studies. Two authors independently screened all retrieved records in two sequential stages. In the first stage, titles and abstracts were screened against the predefined eligibility criteria to exclude obviously irrelevant studies. In the second stage, full texts of potentially eligible studies were retrieved and assessed in detail. Any discrepancies between the two authors at either stage were discussed and, if consensus could not be reached, resolved by a third author.

Inclusion and exclusion criteria

Inclusion criteria

Studies were eligible if they: (1) evaluated the diagnostic performance of an AI model for UGI precancerous lesions using endoscopic images or videos as input data; (2) targeted lesions of interest including esophageal squamous intraepithelial neoplasia (ESIN), Barrett's esophagus with dysplasia, CAG, GIM, or gastric intraepithelial dysplasia; (3) reported sensitivity and specificity, or provided sufficient raw data (true positive (TP), false positive (FP), true negative (TN), false negative (FN)) to calculate these metrics; (4) used histopathological examination as the reference standard, except for one intestinal metaplasia dataset that used a validated endoscopic visual criterion with previously demonstrated concordance with histopathology; and (5) were published in peer-reviewed journals with full text available.

Exclusion criteria

Studies were excluded if they: (1) did not clearly describe the AI model type or failed to report diagnostic performance indicators; (2) focused exclusively on advanced-stage cancers or non-precancerous gastrointestinal conditions; (3) were non-original research, including case reports, narrative reviews, systematic reviews, editorials, or conference abstracts; or (4) used non-endoscopic imaging modalities (e.g., radiological or pathological whole-slide images) as primary input data.

Data extraction

Two reviewers independently extracted data using a prespecified standardized form; discrepancies were resolved through discussion or adjudication by a third reviewer. The following data were extracted: (1) study characteristics: first author, publication year, country, study design (prospective or retrospective), and total sample size; (2) lesion and imaging characteristics: lesion type (ESIN, Barrett's dysplasia, CAG, GIM, or gastric dysplasia),

endoscopic modality (WLE, narrow-band imaging (NBI), magnifying narrow-band imaging (M-NBI), blue laser imaging (BLI), linked color imaging (LCI), or combinations thereof), and whether an endoscopist comparator group was included; (3) AI model characteristics: model architecture (e.g., CNN, deep convolutional neural network, transformer, or hybrid model) and whether the study reported internal or external validation; and (4) diagnostic performance data: TP, FP, TN, FN, sensitivity, specificity, accuracy, and AUC. When data were missing or ambiguous, corresponding authors were contacted for clarification.

Quality assessment

The risk of bias and applicability of included studies were assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2).[17] Risk of bias was assessed across four domains. For patient selection, we assessed whether participants were enrolled consecutively and whether selection bias was present. For the index test, we assessed whether AI model implementation and validation were clearly described. For the reference standard, we assessed whether the reference standard was appropriate for classifying the target condition. For Gong E 2024, dysplasia was confirmed histopathologically, while intestinal metaplasia was assessed using a validated endoscopic visual grading criterion; because the developers of this criterion had previously demonstrated its concordance with histopathology in an independent validation study, this dataset was retained and its reference standard rated as low risk. For flow and timing, we assessed whether the interval between the index test and reference standard was appropriate. Each domain was rated as having a low, high, or unclear risk of bias.

Statistical analysis

When a single study reported diagnostic performance separately for multiple lesion types or imaging modalities, each subgroup was treated as an independent dataset for pooled analysis; accordingly, the 15 included studies contributed a total of 24 datasets.Diagnostic performance data (TP, FP, TN, FN) were pooled using a bivariate random-effects model to simultaneously estimate summary sensitivity and specificity with 95% confidence intervals (CIs). Summary receiver operating characteristic (SROC) curves were constructed and AUC was calculated to reflect overall diagnostic accuracy. All pooled analyses and figure generation were performed using Python (version 3.12) and Meta-DiSc 2.0.[18,19] Pooled sensitivity and specificity were estimated using a DerSimonian–Laird random-effects model applied to logit-transformed proportions[20]; a continuity correction of 0.5 was applied to studies with zero cells.[21] SROC curves and the bivariate random-effects model were implemented according to the method described by Reitsma *et al.*[22] Leave-one-out sensitivity analyses were conducted at both the dataset and study levels following the approach of Viechtbauer and Cheung to assess the influence of individual datasets or studies on the pooled estimates.[23] Datasets from the same study were also aggregated into study-level records and re-pooled as a sensitivity analysis to examine within-study non-independence. Publication bias was assessed using Egger's regression test, Begg's rank correlation test, Deeks' funnel plot asymmetry test, and the Duval and Tweedie trim-and-fill method; a two-sided $P < 0.05$ was considered statistically significant.[24] Subgroup analyses and meta-regression were performed to explore sources of heterogeneity. For subgroup visualization, Figure 1b–f were fitted using the Moses–Littenberg method; the summary points and 95% CIs are the subgroup-specific pooled estimates reported in Table 1. Bivariate meta-regression was performed only for target organ, imaging modality, and study design; analyses by AI architecture and target lesion category were descriptive.

Heterogeneity across studies was quantified using the I^2 statistic and the χ^2 test; $I^2 > 50\%$ was considered indicative of substantial heterogeneity. To explore potential sources of heterogeneity, prespecified subgroup analyses were conducted according to: (1) target organ (esophageal versus gastric precancerous lesions); (2) endoscopic imaging modality (WLE versus image-enhanced endoscopy [IEE], including NBI, M-NBI, BLI, and LCI); (3) study design (retrospective versus prospective); (4) AI model architecture (CNN versus transformer-based models); and (5) target lesion category (intestinal metaplasia, dysplasia, or gastric atrophy, versus precancerous change not further subtyped). Meta-regression was performed for each subgroup variable when the number of datasets per stratum was sufficient.

Results

General information and baseline characteristics of the included studies

The PRISMA 2020 flow diagram (Fig. 2) was generated using the PRISMA2020 R package.[25] A total of 15 studies were included (Fig. 2).[26-40] These studies comprised 24 datasets (as several studies reported results separately for multiple lesion types or imaging modalities) and collectively encompassed 19,578 diagnostic records (Fig. 3a), including image-, video-, and patient-level data, of which 8,336 (42.6%) were positive records and 11,242 (57.4%) were negative records. Publication years ranged from 2020 to 2026, three studies each were published in 2020 (Liu G, Liu X, de Groof AJ), 2021 (Mu G, Lin N, Xu M), 2022 (Wong PK, Gong E [esophageal], Siripoppohn V); four studies were published in 2024 (Tan JL, Li Z, Gong E [gastric], Jhang JY); one study was published in 2023 (Yang J), and one in 2026 (Chiang TH). Studies originated predominantly from China (n = 10), with the remainder from Korea (n = 2), Australia (n = 1), Thailand (n = 1), and the Netherlands (n = 1). Eleven studies addressed gastric precancerous lesions and four targeted esophageal precancerous lesions. Ten studies were retrospective, three were prospective, and two included both retrospective and prospective datasets. At the dataset level, which is the unit used in Table 1, 8 of 24 datasets were prospective and 16 were retrospective, reflecting that Xu M 2021 and Siripoppohn V 2022 each contributed both a prospective and a retrospective dataset. Endoscopic imaging modalities were predominantly WLE (n = 12), with the remainder employing image-enhanced endoscopy techniques. AI architectures ranged from CNN-based models (ResNet-50, TResNet, Two-stream CNN, BiSeNet) to transformer-based and hybrid architectures (BLS²-MSA, GSCNet). Among the 24 datasets, 18 used CNN-based architectures and 6 used transformer-based architectures (Table 1),[26-40] reflecting the diversity of deep learning approaches applied in this field. Detailed baseline characteristics of all included studies are presented in Table 2.

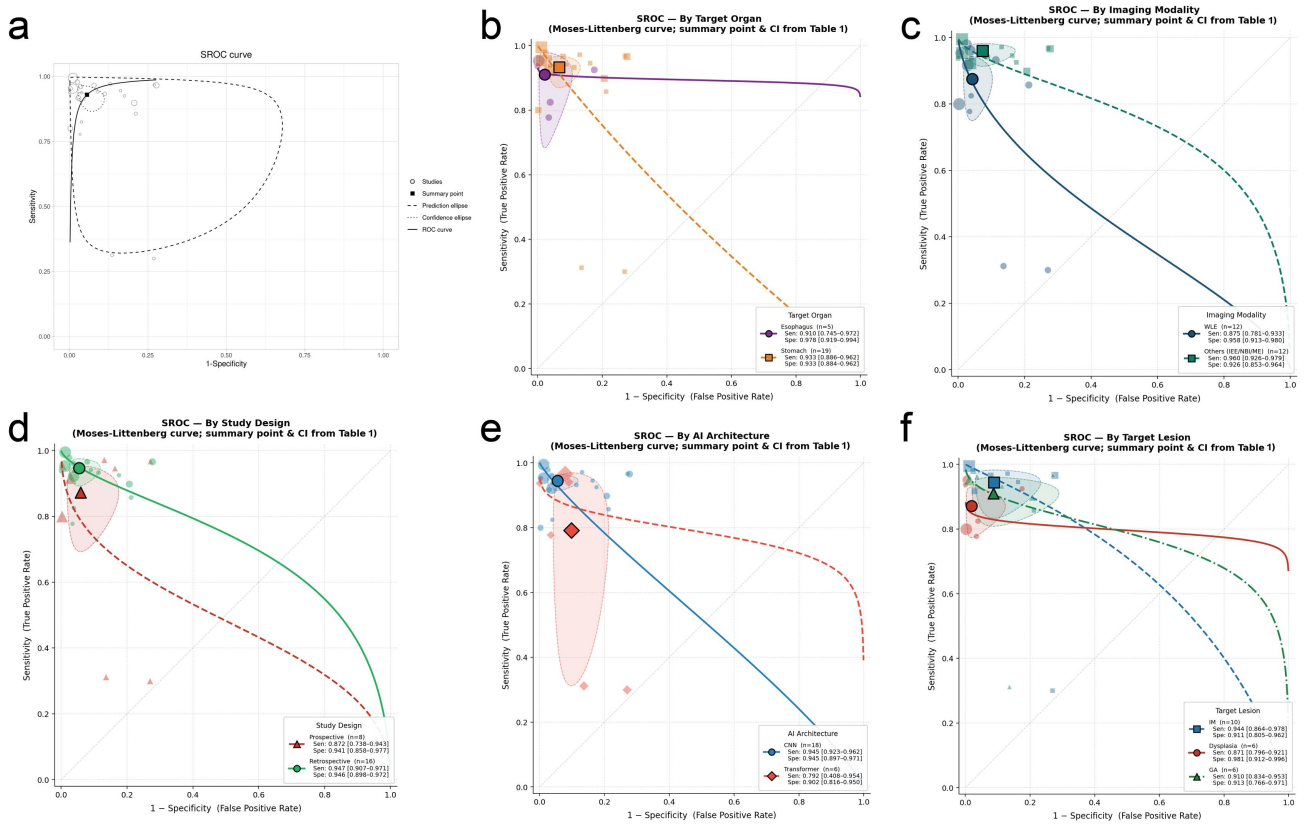

Fig. 1. SROC curves of AI model diagnostic performance. (a) Overall SROC curve under bivariate model analysis. (b–f) Moses–Littenberg subgroup curves; summary points and confidence intervals are from Table 1. Curves in b–f were fitted using the Moses–Littenberg method; whereas the summary points and 95% CIs correspond to the subgroup-specific pooled estimates reported in Table 1; therefore, the summary points may not fall exactly on the plotted curves. Bivariate meta-regression was performed only for target organ, imaging modality, and study design (panels b–d). Comparisons by AI model architecture and target lesion category (panels e and f) are descriptive because meta-regression was not performed for these subgroups. (b) SROC curves comparing the AUC and heterogeneity range for esophageal and gastric lesions. (c) SROC curves comparing the AUC and heterogeneity range for different imaging modalities. (d) SROC curves comparing the AUC and heterogeneity range of prospective vs. retrospective studies. (e) SROC curve by AI model architecture (CNN vs. transformer). (f) SROC curve by target lesion category (intestinal metaplasia, dysplasia, gastric atrophy, precancerous). AI, artificial intelligence; AUC, area under the receiver operating characteristic curve; CI, confidence interval; CNN, convolutional neural network; GA, gastric atrophy; IEE, image-enhanced endoscopy; IM, intestinal metaplasia; ME, magnifying endoscopy; NBI, narrow-band imaging; SROC, summary receiver operating characteristic; WLE, white-light endoscopy.

QUADAS-2 quality assessment

Histopathological examination was used as the reference standard in the included studies, except that intestinal metaplasia in Gong E (2024) was assessed using a validated endoscopic visual criterion previously shown to be concordant with histopathology (see Table 2 footnote) (Table 3).[26-40] The studies encompassed a variety of endoscopic imaging modalities (WLE, NBI, M-NBI, magnifying endoscopy with blue laser imaging, and LCI) and AI architectures (CNN-based, transformer-based, and hybrid models), reflecting the current diversity of deep learning approaches in UGI endoscopy research. Methodological quality was evaluated using the QUADAS-2 tool, which assesses both risk of bias and applicability concerns across four domains: patient selection, index test, reference standard, and flow and timing.

Risk of bias

Patient selection: Of the 15 studies, 5 (33.3%) were classified as low risk, 9 (60.0%) as high risk, and 1 (6.7%) as unclear risk. High-risk ratings in this domain predominantly stemmed from non-consecutive or retrospective image sampling, which may introduce selection bias and potentially overestimate diagnostic accuracy. Index test: All 15 studies (100%) were rated as low risk, reflecting adequate description of AI model implementation and validation procedures. Reference standard: All 15 studies (100%) were rated as low risk; histopathology was used for the target condition in all datasets except the Gong E 2024 intestinal metaplasia dataset, which used the validated visual criterion described above. Flow and timing: All 15 studies (100%) were rated as low risk, indicating appropriate intervals between the index test and reference standard across all included studies.

Table 1. Subgroup analysis results by different factors

Subgroup	n	Sensitivity (95% CI)	Specificity (95% CI)	DOR (95% CI)	LR+ (95% CI)	LR− (95% CI)	FPR (95% CI)	Relative sensitivity (95% CI)	*P* (sensitivity)	Relative specificity (95% CI)	*P* (specificity)	Global test *P*
Target organ												
Esophagus	5	0.91 (0.745–0.972)	0.978 (0.919–0.994)	442.64 (55.17–3551.11)	40.963 (10.462–160.386)	0.093 (0.03–0.287)	0.022 (0.006–0.081)	0.954 (0.907–1.004)	0.127	1.026 (0.912–1.155)	0.636	0.195
Stomach	19	0.933 (0.886–0.962)	0.933 (0.884–0.962)	195.24 (73.13–521.23)	13.936 (7.82–24.834)	0.071 (0.04–0.126)	0.067 (0.038–0.116)	Reference	—	Reference	—	0.195
Imaging modality												
WLE	12	0.875 (0.781–0.933)	0.958 (0.913–0.98)	160.28 (46.43–553.32)	20.857 (9.499–45.799)	0.13 (0.071–0.239)	0.042 (0.02–0.087)	0.911 (0.834–0.996)	0.017	1.035 (0.969–1.105)	0.285	0.005
Others (IEE/NBI/ME)	12	0.96 (0.926–0.979)	0.926 (0.853–0.964)	302.59 (88.66–1032.79)	12.955 (6.28–26.727)	0.043 (0.022–0.083)	0.074 (0.036–0.147)	Reference	—	Reference	—	0.005
Study design												
Prospective	8	0.872 (0.738–0.943)	0.941 (0.858–0.977)	109.54 (24.35–492.81)	14.873 (5.697–38.831)	0.136 (0.062–0.299)	0.059 (0.023–0.142)	Reference	—	Reference	—	0.22
Retrospective	16	0.947 (0.907–0.971)	0.946 (0.898–0.972)	313.02 (109.61–893.89)	17.443 (9.035–33.675)	0.056 (0.031–0.1)	0.054 (0.028–0.102)	1.086 (0.966–1.221)	0.09	1.005 (0.938–1.076)	0.894	0.22
AI architecture												
CNN	18	0.945 (0.923–0.962)	0.945 (0.897–0.971)	295.29 (104.078–837.8)	17.062 (8.984–32.949)	0.058 (0.039–0.086)	0.055 (0.029–0.103)	Not performed	—	Not performed	—	Not performed
Transformer	6	0.792 (0.408–0.954)	0.902 (0.816–0.950)	34.99 (3.058–400.268)	8.074 (2.218–19.169)	0.231 (0.048–0.725)	0.098 (0.050–0.184)	Not performed	—	Not performed	—	Not performed
Target lesion												
IM	10	0.944 (0.864–0.978)	0.911 (0.805–0.962)	171.56 (32.30–911.19)	10.633 (4.442–25.870)	0.062 (0.023–0.169)	0.089 (0.038–0.195)	Not performed	—	Not performed	—	Not performed
Dysplasia	6	0.871	0.981	336.72	45.647	0.132	0.019	Not	—	Not	—	Not

Table 1. Subgroup analysis results by different factors (*continued*)

Subgroup	n	Sensitivity (95% CI)	Specificity (95% CI)	DOR (95% CI)	LR+ (95% CI)	LR− (95% CI)	FPR (95% CI)	Relative sensitivity (95% CI)	*P* (sensitivity)	Relative specificity (95% CI)	*P* (specificity)	Global test *P*
		(0.796–0.921)	(0.912–0.996)	(81.23–1395.80)	(9.029–236.298)	(0.079–0.224)	(0.004–0.088)	performed		performed		performed
GA	6	0.910 (0.834–0.953)	0.913 (0.766–0.971)	103.21 (14.98–711.22)	10.430 (3.568–32.764)	0.098 (0.048–0.216)	0.087 (0.029–0.234)	Not performed	—	Not performed	—	Not performed
Precancerous	2							Not performed	—	Not performed	—	Not performed

n = number of datasets (24 total in every subgroup variable). Relative sensitivity/specificity and *P* values are from bivariate meta-regression, with stomach, other imaging modalities (IEE/NBI/ME), and prospective study design as the reference categories, respectively; global test *P* reflects the overall significance of the subgroup variable as a covariate. Meta-regression was not performed for AI architecture because of the small number of transformer-based datasets (*n* = 6); results are descriptive only. Meta-regression was not performed for target lesion because of the small number of precancerous datasets (*n* = 2); results are descriptive only. Precancerous subgroup pooled estimates are not reported given *n* = 2. AI, artificial intelligence; CI, confidence interval; CNN, convolutional neural network; DOR, diagnostic odds ratio; FPR, false-positive rate; GA, gastric atrophy; IEE, image-enhanced endoscopy; IM, intestinal metaplasia; LR+, positive likelihood ratio; LR−, negative likelihood ratio; ME, magnifying endoscopy; NBI, narrow-band imaging; WLE, white-light endoscopy.

Applicability concerns

Patient selection: Five studies (33.3%) demonstrated low applicability concerns, 9 (60.0%) had high applicability concerns, and 1 (6.7%) was rated unclear, mirroring the risk of bias pattern and reflecting the predominance of retrospective, single-center designs that may limit generalizability to routine clinical practice. Index test: All 15 studies (100%) had low applicability concerns for the index test, indicating that the AI models and imaging protocols were considered applicable to the review question. Reference standard: All 15 studies (100%) were at low applicability concern; histopathology was used in all datasets except the Gong E 2024 intestinal metaplasia dataset, for which the validated visual criterion described above was used.

Results of univariate model analysis

Overall diagnostic performance

Based on the univariate random-effects model, the pooled sensitivity of the AI model in the diagnosis of upper gastrointestinal precancerous lesions was 0.928 (95% CI: 0.88–0.96), the specificity was 0.94 (95% CI: 0.91–0.97), and the diagnostic odds ratio (DOR) was 220.47 (95% CI: 100.86–481.89) (Table 4). The positive likelihood ratio (LR+) was 16.75 (95% CI: 9.78–28.68), and the negative likelihood ratio was 0.076 (95% CI: 0.046–0.125), indicating high discriminatory ability of AI-assisted endoscopy in the detection of upper gastrointestinal precancerous lesions.

Sensitivity and specificity analysis

The heterogeneity in sensitivity was 91.6%, indicating extremely high heterogeneity among the included studies ($P < 0.01$). Sensitivity across datasets ranged from 0.30 to 0.99. Notably, studies targeting intestinal metaplasia (IM) with WLE tended to report lower sensitivity estimates (e.g., Chiang TH 2026 IM subgroup, sensitivity 0.30, 95% CI: 0.15–0.49), which may reflect the inherently subtle endoscopic appearance of IM under standard imaging.

Heterogeneity in specificity was equally pronounced, with an I^2 of 96.7% and specificities ranging from 0.73 to 1.00 across individual datasets. In the specificity forest plot (Fig. 3c), the summary estimate was 0.94 (95% CI: 0.91–0.97). At the subgroup level, specificity varied by target lesion category: dysplasia datasets achieved the highest pooled specificity (0.981), whereas intestinal metaplasia and gastric atrophy datasets were lower and more variable (0.911 and 0.913, respectively; Table 1). Accordingly, in the meta-regression the significant association between imaging modality and performance was confined to sensitivity (0.960 vs. 0.875; $P = 0.017$), with no corresponding difference in specificity (relative specificity 1.035; $P = 0.285$). Figure 3d presents the overall SROC curve based on the bivariate model (the same curve shown in Fig. 1a), with a summary AUC of 0.9661, indicating excellent overall discriminative ability (Fig. 3d).

Results of bivariate model analysis

Based on the bivariate random-effects model, the summary sensitivity and specificity of the AI model in the diagnosis of upper gastrointestinal precancerous lesions were 0.93 (95% CI:

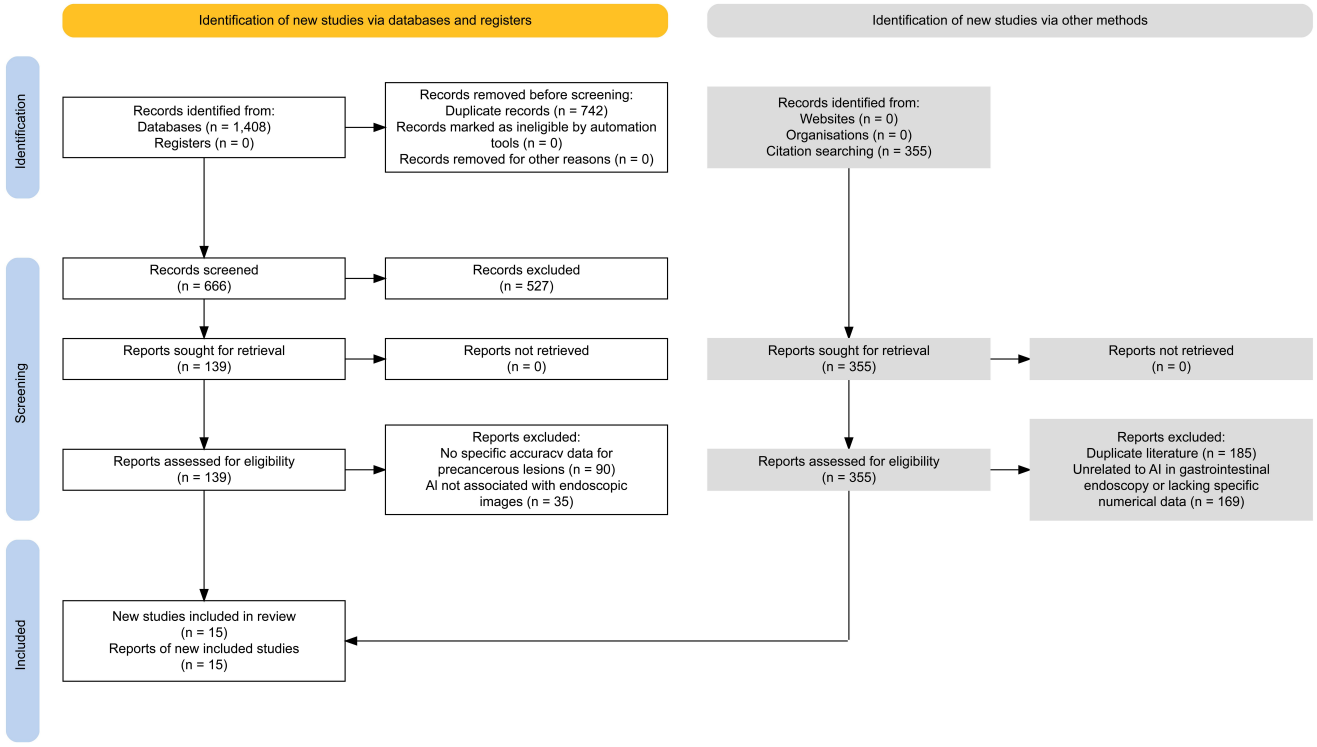

Fig. 2. PRISMA 2020 flow diagram.[25] AI, artificial intelligence; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

0.886–0.958) and 0.944 (95% CI: 0.906–0.967), respectively, and the DOR was 223.79 (95% CI: 92.24–542.94), indicating that AI-assisted endoscopy has a high discriminatory ability between precancerous and non-precancerous lesions. The LR+ was 16.66 (95% CI: 9.69–28.65), indicating that a positive AI result substantially increases the post-test probability of a precancerous lesion; the negative likelihood ratio was 0.074 (95% CI: 0.045–0.124), indicating that a negative AI result substantially lowers the post-test probability of precancerous lesions. The false-positive rate was 0.056 (95% CI: 0.033–0.094) (Table 4).

The bivariate I^2 was 0.73 (73%), indicating substantial heterogeneity across the included datasets ($P < 0.01$). The sensitivity forest plot demonstrated considerable variability across studies, with individual sensitivities ranging from 0.30 to 0.99 (Fig. 4a), while specificities ranged from 0.72 to 1.00 (Fig. 4b). Studies employing image-enhanced endoscopy modalities such as magnifying narrow-band imaging and magnifying endoscopy with blue laser imaging tended to show higher sensitivity, whereas studies focusing on intestinal metaplasia under white-light endoscopy showed comparatively lower sensitivity, likely reflecting the subtle endoscopic appearance of IM. The ROC plane scatter plot (Fig. 4c) demonstrated that the majority of studies clustered in the upper-left region, consistent with high overall diagnostic accuracy. The 95% prediction ellipse was broad, reflecting the expected range of diagnostic performance in future comparable studies.

ROC and model consistency

The SROC curve showed that the summary AUC was 0.9661, indicating excellent overall discriminative ability (Fig. 1a). The majority of individual study points clustered in the upper-left region of the ROC plane, indicating the coexistence of high sensitivity and high specificity across most included datasets. The 95% confidence ellipse was narrow and centered near the summary point, reflecting reliable estimation of the pooled diagnostic accuracy. However, the 95% prediction ellipse was substantially wider, extending toward the lower-right region of the ROC plane, suggesting considerable variability in the expected diagnostic performance of AI models in future comparable studies. Subgroup-stratified Moses–Littenberg SROC plots (Fig. 1b–f; summary points and confidence intervals from Table 1) further illustrated the sources of this heterogeneity. When stratified by target organ (Fig. 1b), esophageal studies (purple summary point) and gastric studies (orange summary point) showed distinct distributional patterns, with esophageal studies tending to cluster closer to the upper-left corner, potentially reflecting the more distinct endoscopic features of esophageal dysplasia compared with gastric precancerous lesions. Stratification by imaging modality (Fig. 1c) revealed that studies using WLE (navy circles) demonstrated a wider spread of diagnostic performance compared with those employing other modalities (teal squares), consistent with the known limitations of WLE in detecting subtle mucosal changes. Stratification by study design (Fig. 1d) showed that prospective studies (triangles) and retrospective studies (circles) were similarly distributed across the ROC plane, suggesting that study design alone did not fully account for the observed heterogeneity. Stratification by AI model architecture (Fig. 1e) showed that CNN-based models (blue circle) achieved a tightly clustered summary point close to the upper-left

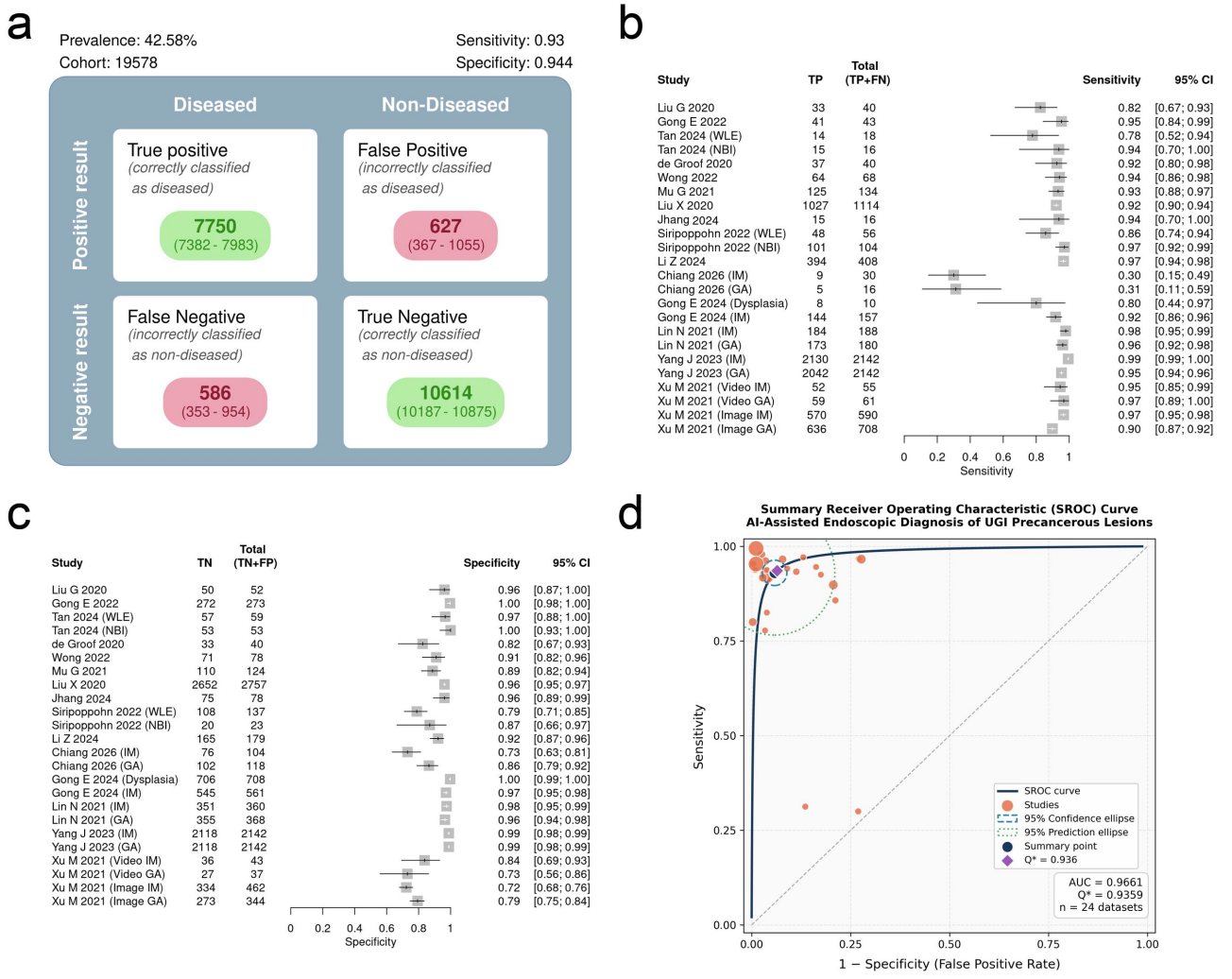

Fig. 3. Univariate model analysis results of AI model diagnostic performance. (a) Summary result of pooled data. Panel a shows an illustrative distribution of diagnostic outcomes (TP, FP, FN, TN) across 19,578 diagnostic records using the aggregate positive-record proportion (42.6%), computed from the pooled model sensitivity and specificity; it is not a direct sum of the raw per-dataset counts and does not represent patient-level prevalence. Raw per-dataset counts are provided in Supplementary File 3. (b) Forest plot of sensitivity. (c) Forest plot of specificity. (d) SROC curve. Q*, the point on the SROC curve at which sensitivity equals specificity. AI, artificial intelligence; AUC, area under the receiver operating characteristic curve; CI, confidence interval; FN, false negative; FP, false positive; GA, gastric atrophy; IM, intestinal metaplasia; NBI, narrow-band imaging; SROC, summary receiver operating characteristic; TN, true negative; TP, true positive; UGI, upper gastrointestinal; WLE, white-light endoscopy.

corner with a comparatively narrow prediction ellipse, whereas transformer-based models (red diamond) showed a summary point shifted toward lower sensitivity and a markedly wider prediction ellipse extending toward the lower-right region of the ROC plane, reflecting the smaller number and greater architectural/lesion-type heterogeneity of the transformer-based datasets. Stratification by target lesion category (Fig. 1f) showed that dysplasia (red circle) achieved the summary point closest to the upper-left corner, consistent with its highest specificity among the pooled lesion categories, while intestinal metaplasia (blue square) and gastric atrophy (green triangle) showed broadly overlapping distributions with wider, similarly shaped prediction ellipses, indicating that these two lesion types were not clearly separated in overall diagnostic performance.

Subgroup analysis results

Subgroup analyses were performed according to target organ, imaging modality, study design, AI model architecture, and target lesion category, with results summarized in Table 1 and illustrated by Moses–Littenberg SROC curves in Figure 1b–f.

By target organ, esophageal studies (n = 5 datasets) showed numerically higher specificity than gastric studies (0.978, 95% CI: 0.919–0.994 vs. 0.933, 95% CI: 0.884–0.962), alongside a substantially higher DOR (442.642, 95% CI: 55.175–3551.110 vs. 195.238, 95% CI: 73.130–521.231) and LR+ (40.963 vs. 13.936).

 DOI: 10.14218/CSP.2026.00010 | 2026

Table 2. Baseline characteristics of the included studies

Study	Year	Country	Study design	Target organ	Target lesion	Imaging type	AI model	Reference standard	Test sample size[d]
Liu G[26]	2020	China	Retrospective	Esophagus	Dysplasia	WLE	Two-stream CNN	Pathology	129 images
Wong PK[27]	2022	China	Retrospective	Stomach	IM	M-NBI	BLS^2-MSA	Pathology	146 images
Gong E[28]	2022	Korea	Prospective	Esophagus	Dysplasia	WLE	Neuro-T	Pathology	836 images
Mu G[29]	2021	China	Retrospective	Stomach	Gastric atrophy	WLE	UNet++ + ResNet-50	Pathology	258 images / 80 videos
Liu X[30]	2020	China	Retrospective	Stomach	Precancerous (LGN)	M-NBI	ResNet-50	Pathology	3,871 images
Jhang JY[31]	2024	China	Retrospective	Stomach	Precancerous (CGI)	WLE	GSCNet	Pathology	912 images (304 patients)
Tan JL[32]	2024	Australia	Retrospective	Esophagus	Dysplasia	WLE / NBI	ViT / DeiT	Pathology	146 images (total)
Siripoppohn V[33a]	2022	Thailand	Prospective/retrospective	Stomach	IM	WLE / NBI	BiSeNet	Pathology	320 images (total)
Li Z[34]	2024	China	Retrospective	Stomach	IM	IEE (NBI)	Deformable transformer	Pathology	587 images
Chiang TH[35]	2026	China	Prospective	Stomach	IM / GA	WLE	Vision transformer	Pathology	134 patients
Gong E[36b]	2024	Korea	Prospective	Stomach	Dysplasia / IM	WLE	Neuro-X (CNN) / Edge AI	Pathology / validated visual criterion for IM	718 images
Lin N[37]	2021	China	Retrospective	Stomach	IM / GA	WLE	TResNet	Pathology	548 images (273 patients)
Yang J[38]	2023	China	Retrospective	Stomach	IM / GA	WLI + LCI	LAG	Pathology	4,284 images
Xu M[39c]	2021	China	Prospective/retrospective	Stomach	IM / GA	ME-NBI/BLI	CNN	Pathology	1,052 images ×2 (IM/GA) / 98 videos ×2 (IM/GA)
de Groof AJ[40]	2020	Netherlands	Retrospective	Esophagus	Neoplasia / dysplasia	WLE + NBI	CNN	Pathology	80 images

[a]Siripoppohn 2022 contributed two datasets analyzed separately: a retrospective white-light (WLE) image set and a prospective narrow-band imaging (NBI) validation set; these are treated as separate design categories in Table 1 and all dataset-level analyses. [b]Gong E 2024: dysplasia was confirmed histopathologically; intestinal metaplasia (IM) was graded using a validated endoscopic visual criterion previously shown by the same authors to agree with histopathology, and was therefore retained as meeting the review's histopathology-based inclusion criterion (see Quality Assessment). [c]Xu M 2021 contributed four datasets analyzed separately by acquisition setting: two retrospective, static-image datasets (Image IM, Image GA; 5,198 ME-NBI/BLI images from 5 hospitals used for model development, internal and external testing) and two prospective, real-time

video datasets (Video IM, Video GA; 98 video clips from 77 consecutively enrolled patients at Wuhan University People's Hospital, Jan–Oct 2020, used for real-time CADe validation). Because the prospective and retrospective datasets from this study share a common development pipeline, their subgroup estimates are not statistically independent; this is discussed as a limitation. [d]Test Sample Size refers to the test/validation sample reported in the original study and may differ from the diagnostic observations included in the meta-analysis; the exact TP, FP, TN, and FN counts used for pooling are provided in Supplementary File 3. AI, artificial intelligence; BiSeNet, Bilateral Segmentation Network; BLI, blue laser imaging; BLS²-MSA, Broad Learning System (stacking) with Multi-Scale Attention; CADe, computer-aided detection; CGI, corpus-predominant gastritis index; CNN, convolutional neural network; DeiT, data-efficient image Transformer; FN, false negative; FP, false positive; GA, gastric atrophy; GSCNet, Gastric Section Correlation Network; IEE, image-enhanced endoscopy; LAG, Local Attention Grouping; LCI, linked color imaging; LGN, low-grade neoplasia; M-NBI, magnifying narrow-band imaging; ME-BLI, magnifying endoscopy with blue laser imaging; ME-NBI, magnifying endoscopy with narrow-band imaging; NBI, narrow-band imaging; Neuro-T, automated deep learning model-training software; TN, true negative; TP, true positive; TResNet, GPU-dedicated convolutional neural network architecture; ViT, Vision Transformer; WLI, white-light imaging.

However, meta-regression analysis revealed no statistically significant difference in sensitivity (relative sensitivity: 0.954, 95% CI: 0.907–1.004, $P = 0.127$) or specificity (relative specificity: 1.026, 95% CI: 0.912–1.155, $P = 0.636$) between gastric and esophageal studies, and the global test was non-significant ($P = 0.195$), indicating that target organ did not significantly contribute to the observed heterogeneity.

By imaging modality, studies employing image-enhanced endoscopy or other advanced modalities (IEE/NBI/ME, $n = 12$ datasets) achieved higher sensitivity than WLE studies (0.960, 95% CI: 0.926–0.979 vs. 0.875, 95% CI: 0.781–0.933). Meta-regression confirmed that WLE was associated with significantly lower sensitivity compared with other modalities (relative sensitivity: 0.911, 95% CI: 0.834–0.996, $P = 0.017$). Although the difference in specificity did not reach statistical significance (relative specificity: 1.035, 95% CI: 0.969–1.105, $P = 0.285$), the global test demonstrated a statistically significant overall difference between modality subgroups ($P = 0.005$), confirming that imaging modality was a significant source of heterogeneity in the dataset.

By study design, retrospective studies ($n = 16$ datasets) showed numerically higher sensitivity (0.947, 95% CI: 0.907–0.971 vs. 0.872, 95% CI: 0.738–0.943) and comparable specificity (0.946 vs. 0.941) relative to prospective studies ($n = 8$ datasets). However, meta-regression analysis showed no significant difference in sensitivity (relative sensitivity: 1.086, 95% CI: 0.966–1.221, $P = 0.09$) or specificity (relative specificity: 1.005, 95% CI: 0.938–1.076, $P = 0.894$) between retrospective and prospective studies, with a non-significant global test ($P = 0.22$), suggesting that study design did not independently explain the heterogeneity observed.

By AI model architecture, CNN-based models ($n = 18$ datasets) achieved a pooled sensitivity of 0.945 (95% CI: 0.923–0.962) and specificity of 0.945 (95% CI: 0.897–0.971), compared with 0.792 (95% CI: 0.408–0.954) and 0.902 (95% CI: 0.816–0.950) for transformer-based models ($n = 6$ datasets). The markedly wider confidence interval for the transformer subgroup reflects both the small number of transformer-based datasets and their heterogeneous composition (spanning Barrett's dysplasia, gastric intestinal metaplasia, and gastric atrophy); this comparison should therefore be interpreted as preliminary rather than a definitive architecture-level ranking.

By target lesion category, sensitivity and specificity varied across lesion types: intestinal metaplasia (IM, $n = 10$ datasets) showed sensitivity of 0.944 (95% CI: 0.864–0.978) and specificity of 0.911 (95% CI: 0.805–0.962); dysplasia ($n = 6$ datasets) showed the highest specificity at 0.981 (95% CI: 0.912–0.996) but a comparatively lower sensitivity of 0.871 (95% CI: 0.796–0.921); and gastric atrophy ($n = 6$ datasets) showed intermediate performance (sensitivity 0.910, 95% CI: 0.834–0.953; specificity 0.913, 95% CI: 0.766–0.971). The remaining precancerous-lesion category (Liu X 2020 and Jhang JY 2024, $n = 2$ datasets) was considered too small to support a stable pooled estimate and is reported descriptively only (Table 1). These lesion-specific differences indicate that the overall pooled estimate should not be treated as uniformly applicable across lesion types, and that AI performance is better characterized at the level of individual lesion categories than by a single pooled figure.

In summary, among the subgroup variables formally evaluated by meta-regression, imaging modality was significantly associated with heterogeneity in AI diagnostic performance (global $P = 0.005$), with WLE demonstrating significantly lower sensitivity compared with image-enhanced modalities. Neither target organ nor study design reached statistical significance in meta-regression analyses, and meta-regression was not extended to the AI-architecture or target-lesion subgroups because of the small number of datasets in several strata (e.g., transformer, $n = 6$; precancerous, $n = 2$); these subgroup comparisons are therefore descriptive rather than confirmatory.

Publication bias assessment

Publication bias was assessed using Egger's regression test, Begg's rank correlation test, Deeks' funnel plot asymmetry test, and the trim-and-fill method. Egger's test showed no statistically significant small-study effect (intercept = −0.152, standard error = 0.714, $P = 0.833$). Begg's rank correlation test similarly indicated no significant asymmetry (Kendall's $\tau = 0.073$, $P = 0.641$). Deeks' funnel plot asymmetry test, which is specifically recommended for diagnostic accuracy meta-analyses, yielded a P value of 0.053, marginally above the threshold for statistical significance. Visual inspection of the funnel plot (Fig. 5) showed an approximately symmetric distribution of study points around the pooled DOR estimate. The trim-and-fill

 DOI: 10.14218/CSP.2026.00010 | 2026

Table 3. QUADAS-2 quality assessment results of included studies

Study	Year	Risk of bias				Applicability		
		Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Liu G[26]	2020	High	Low	Low	Low	High	Low	Low
Wong PK[27]	2022	High	Low	Low	Low	High	Low	Low
Gong E[28]	2022	Low	Low	Low	Low	Low	Low	Low
Mu G[29]	2021	Unclear	Low	Low	Low	Unclear	Low	Low
Liu X[30]	2020	High	Low	Low	Low	High	Low	Low
Jhang JY[31]	2024	High	Low	Low	Low	High	Low	Low
Tan JL[32]	2024	High	Low	Low	Low	High	Low	Low
Siripoppohn V[33]	2022	Low	Low	Low	Low	Low	Low	Low
Li Z[34]	2024	High	Low	Low	Low	High	Low	Low
Chiang TH[35]	2026	Low	Low	Low	Low	Low	Low	Low
Gong E[36]	2024	Low	Low	Low	Low	Low	Low	Low
Lin N[37]	2021	High	Low	Low	Low	High	Low	Low
Yang J[38]	2023	High	Low	Low	Low	High	Low	Low
Xu M[39]	2021	Low	Low	Low	Low	Low	Low	Low
de Groof AJ[40]	2020	High	Low	Low	Low	High	Low	Low

QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies-2.

Table 4. Summary results of AI model diagnostic performance

Index	Estimate	95% LCI	95% UCI
Univariate random-effects model			
Sensitivity	0.928	0.883	0.957
Specificity	0.945	0.906	0.968
DOR	220.47	100.86	481.89
LR+	16.75	9.78	28.68
LR−	0.076	0.046	0.125
FPR	0.055	0.032	0.094
Bivariate random-effects model			
Sensitivity	0.930	0.886	0.958
Specificity	0.944	0.906	0.967
DOR	223.79	92.24	542.94
LR+	16.66	9.69	28.65
LR−	0.074	0.045	0.124
FPR	0.056	0.033	0.094

AI, artificial intelligence; DOR, diagnostic odds ratio; FPR, false-positive rate; LCI, lower confidence interval; LR+, positive likelihood ratio; LR−, negative likelihood ratio; UCI, upper confidence interval.

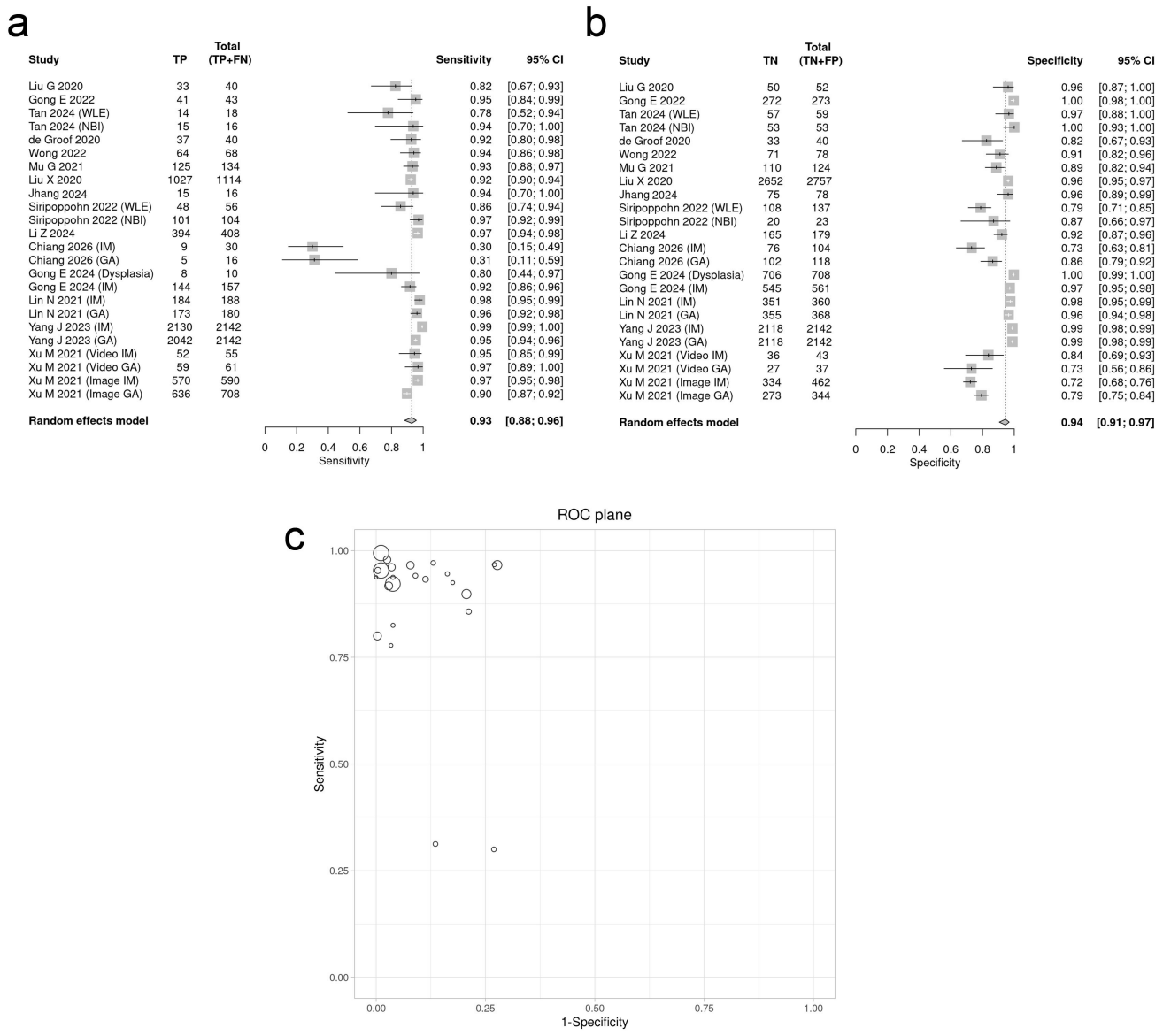

Fig. 4. Bivariate model analysis results of AI model diagnostic performance. (a) Forest plot of sensitivity. (b) Forest plot of specificity. (c) ROC plane. AI, artificial intelligence; CI, confidence interval; FN, false negative; FP, false positive; GA, gastric atrophy; IM, intestinal metaplasia; NBI, narrow-band imaging; ROC, receiver operating characteristic; TN, true negative; TP, true positive; WLE, white-light endoscopy.

analysis (Duval & Tweedie's L0 estimator, applied in Python to the same random-effects log(DOR) model used in the leave-one-out analysis above) identified 2 potentially missing studies, yielding an adjusted pooled DOR of 136.10 (95% CI: 55.64–332.86), compared with an unadjusted DOR of 186.98 (95% CI: 80.55–434.03) from the same univariate log(DOR) model reported as the leave-one-out baseline above; this value is numerically distinct from, and not a substitute for, the bivariate DOR of 223.79 reported as this study's primary effect estimate (Table 4), with the adjusted estimate remaining statistically significant. These results indicate no substantial evidence of publication bias, and the overall conclusions of this meta-analysis are unlikely to be materially affected by selective publication.

Leave-one-out sensitivity analysis

Leave-one-out sensitivity analysis was performed to evaluate the robustness of the pooled estimates, using a univariate DerSimonian–Laird random-effects model implemented in Python and applied directly to each dataset's sensitivity, specificity, and DOR. Because this implementation differs both from the bivariate model underlying the study's primary results and from the MetaDiSc univariate summary reported in Table 4, its baseline point estimates (sensitivity 0.923, specificity 0.938) are close to, but not identical to, those values (e.g., the Table 4 univariate sensitivity of

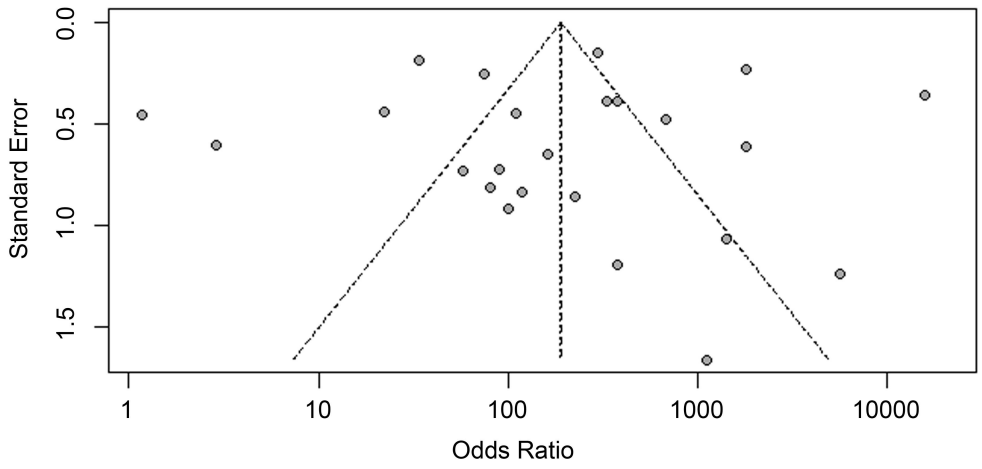

Fig. 5. Funnel plot for publication bias assessment of AI-assisted endoscopic diagnosis of upper gastrointestinal precancerous lesions. AI, artificial intelligence.

0.928); this reflects only the different pooling implementation, not any difference in the underlying data. Sequential exclusion of each of the 24 datasets yielded a pooled sensitivity ranging from 0.913 to 0.934 and specificity ranging from 0.931 to 0.943, with no single dataset altering the pooled sensitivity or specificity by more than 0.02 from the overall estimate (baseline: sensitivity = 0.923, 95% CI: 0.888–0.948; specificity = 0.938, 95% CI: 0.897–0.964; DOR = 186.98, 95% CI: 80.55–434.03), confirming that our conclusions are not driven by any individual study. Exclusion of the two Chiang 2026 datasets, which reported the lowest sensitivity among included studies (0.30–0.31), resulted in only a marginal increase in pooled sensitivity (+0.009 to +0.011), while exclusion of Yang J 2023 (IM), the largest dataset by sample size, produced the largest single-dataset effect (ΔSensitivity = −0.010, ΔSpecificity = −0.006) due to its high statistical weight. Full leave-one-out results for all 24 datasets are provided in Supplementary File 3.

Discussion

The present systematic review and meta-analysis, incorporating 15 studies and 24 datasets comprising 19,578 diagnostic records, indicates that AI-assisted endoscopy has high diagnostic accuracy for upper gastrointestinal precancerous lesions, with a pooled sensitivity of 0.930 (95% CI: 0.886–0.958), specificity of 0.944 (95% CI: 0.906–0.967), DOR of 223.79, and AUC of 0.9661. The positive likelihood ratio of 16.658 and negative likelihood ratio of 0.074 indicate that AI-assisted endoscopy has meaningful discriminatory ability, substantially increasing the post-test probability when positive and lowering it when negative. Among the variables formally evaluated by meta-regression, imaging modality was associated with heterogeneity (global $P = 0.005$), with image-enhanced endoscopy (IEE/NBI/ME) achieving markedly higher sensitivity than white-light endoscopy (0.960 vs. 0.875). No evidence of substantial publication bias was detected across four complementary assessments. These findings support and extend prior evidence.[13,41] This review incorporates a broader spectrum of lesion types (ESIN, Barrett's dysplasia, CAG, GIM, and gastric dysplasia), as well as literature published through March 2026, and provides a broad synthesis of AI diagnostic performance across the UGI precancerous lesion spectrum.

The pooled diagnostic estimates observed in the present analysis are broadly consistent with, yet extend beyond, those reported in prior meta-analyses of AI for individual UGI lesion categories. Dilaghi *et al.*[13] reported a pooled diagnostic accuracy of 90% for AI-based detection of gastric precancerous lesions (and 80% for *H. pylori* infection), whereas the present analysis, which spans a wider lesion spectrum, reports sensitivity and specificity directly and incorporates more recent architectures. The higher sensitivity observed with image-enhanced endoscopy modalities than with WLE (sensitivity 0.960 vs. 0.875; $P = 0.017$) is consistent with findings from individual high-quality studies and suggests potential complementarity between AI and advanced imaging. Yuan *et al.*[42] evaluated an AI-assisted system for detecting esophageal precancerous lesions under WLE and non-magnified NBI in a multicenter randomized controlled trial. Similarly, Xu *et al.*[39] reported a multicenter computer-aided detection system for gastric atrophy and GIM using magnifying narrow-band imaging and blue laser imaging, achieving accuracy of 0.901 in internal and 0.864 in external validation. These data suggest that the diagnostic yield of AI may be enhanced when deployed in conjunction with advanced endoscopic imaging, and that imaging modality should be considered an important factor associated with AI system performance in both study design and clinical implementation.

The rapid advancement of AI endoscopy represents a timely opportunity to address the persistent gaps in UGI cancer surveillance, particularly given the growing global burden of gastric and esophageal cancers and the well-documented limitations of conventional WLE-based surveillance. As health systems increasingly pursue population-level endoscopic screening programs, particularly in high-incidence regions of East Asia, AI-assisted endoscopy may help standardize detection quality, reduce interobserver variability, and extend specialist-

level diagnostic capability to lower-resource settings. However, translating the high accuracy observed in controlled research settings into routine clinical benefit will require several critical developments.[43] Future studies should prioritize prospective, multicenter designs with consecutive patient enrollment to minimize spectrum bias and enhance generalizability. Standardized reporting of performance metrics stratified by image quality, lesion subtype, and endoscopist experience level would facilitate more meaningful cross-study comparisons. Furthermore, the development of interpretable AI architectures capable of providing lesion-level visual explanations alongside diagnostic outputs may help build clinician trust and support regulatory review. The integration of AI with real-time video endoscopy, rather than static image analysis, represents a natural next step, as evidenced by emerging systems providing frame-by-frame lesion delineation during live procedures.[42] Addressing these priorities will be important for bridging the gap between experimental research and clinical-grade AI tools for UGI precancerous lesion surveillance.

An important caveat concerns the direction and likely magnitude of the patient-selection bias identified above. Because 9 of 15 studies (60%) drew their test images from retrospectively curated, non-consecutive samples, often selecting technically adequate, unambiguous images for algorithm testing, the sensitivity and specificity estimates reported here are best interpreted as an upper bound on achievable performance under optimal, curated conditions, rather than a reliable estimate of accuracy in unselected, real-world screening populations, where image quality is more variable and lesion presentations are less clear-cut. Prospective studies with consecutive enrollment, reflecting the full spectrum of lesion appearances encountered in routine practice, are needed to determine the extent to which current estimates will generalize to unselected screening populations.

The clinical value of AI-assisted diagnosis is also likely to vary substantially across intended-use scenarios, and a one-size-fits-all interpretation of these pooled estimates is unlikely to be appropriate. AI in endoscopy can serve several distinct roles, including flagging suspicious areas during screening, guiding targeted biopsy sampling, and informing treatment planning; its clinical utility is not equivalent across these roles. Because the included studies overwhelmingly evaluated AI as a static-image or video-clip classifier rather than in these more specific decision-support capacities, the available evidence most directly supports AI's use in flagging lesions for endoscopist review during screening, a role in which randomized evidence has shown that AI assistance can improve endoscopists' diagnostic performance for Barrett's neoplasia.[44] Its incremental value is likely to be considerably smaller in scenarios where lesion size or morphology already mandates endoscopic resection irrespective of the AI-assigned grade, since management would not change regardless of the algorithm's output. Clarifying the specific clinical scenarios in which AI is most likely to add value, and those in which its contribution may be limited, will help clinicians and readers critically evaluate the applicability of these findings to their own practice setting.

A further limitation relates to the gap between offline, static-image (or preselected video-clip) validation, the paradigm used by nearly all included studies, and the demands of real-time video endoscopy in clinical practice. Diagnostic performance measured on curated images does not necessarily translate to frame-by-frame performance during continuously moving, live endoscopic examinations, where motion artifact, variable insufflation, mucus, and bubbles introduce sources of difficulty that static-image benchmarks do not capture. Findings regarding the clinical readiness of AI-assisted UGI endoscopy should therefore be extrapolated to routine practice with caution until performance has been confirmed in prospective, real-time video studies.

Several limitations of the present meta-analysis warrant consideration. First, substantial heterogeneity was observed across included studies (bivariate $I^2 = 73\%$), which, although partially explained by imaging modality, could not be fully accounted for by the prespecified subgroup variables, suggesting the presence of additional unmeasured sources of variability such as lesion prevalence, AI model architecture, and training data composition. Notably, diagnostic accuracy also varied meaningfully by target lesion category (Table 1; Fig. 1f), with dysplasia showing the highest specificity but comparatively lower sensitivity than intestinal metaplasia or gastric atrophy; pooling across lesion types into a single overall estimate, as in the headline result, therefore risks obscuring clinically important differences in AI performance by lesion subtype, and lesion-specific estimates should be preferred when applying these results to a specific clinical question. Second, among the included studies, 9 of 15 studies (60%) were judged to be at high risk of bias in the patient selection domain of QUADAS-2, primarily owing to retrospective image curation and non-consecutive enrollment; the performance estimates derived from such studies are likely to overestimate real-world diagnostic accuracy. Third, the marked geographic concentration of included studies, with 10 of 15 (66.7%) originating from China, limits the generalizability of findings to populations with different disease epidemiology, *H. pylori* prevalence, and endoscopic practice patterns. The GRAIDS study, the largest multicenter AI endoscopy study to date, demonstrated that performance equivalent to expert endoscopists could only be achieved after training on images from multiple centers,[41] underscoring that single-center models are unlikely to generalize without geographically diverse training data. Federated learning approaches, which enable cross-institutional model training without centralizing patient data, represent a promising methodological avenue to address this challenge.[45] Fourth, multiple datasets derived from the same study and, in some cases, the same patient cohort were treated as independent observations in the primary analysis. To assess the impact of this non-independence, we performed leave-one-out sensitivity analyses at both the dataset and study level, which showed that excluding any single dataset or study changed the pooled sensitivity or specificity by no more than 2 percentage points, indicating that the overall results were not driven by any individual study. We additionally collapsed the 24 datasets into 15 independent study-level records (summing counts within each study) and re-pooled the estimates as a study-level aggregated sensitivity analysis: the point estimates were essentially unchanged (sensitivity 0.919 vs. 0.924; specificity 0.945 vs. 0.938), but the corresponding 95% confidence intervals widened by approximately 21–26% relative to the primary 24-dataset analysis. This confirms that treating non-

independent datasets as independent produced somewhat anti-conservative confidence intervals, although it did not materially alter the point estimates or overall conclusions; subgroup and meta-regression *P* values in this study should nonetheless be interpreted with appropriate caution (Supplementary File 3). In addition, the included datasets used different units of analysis, including images, video clips, and patients; pooled estimates combining these analytic units should therefore be interpreted with caution. Notwithstanding these limitations, the consistency of findings across diverse study designs, lesion types, and geographic settings supports the robustness of the overall conclusions.

Conclusions

This systematic review and meta-analysis indicated that AI-assisted endoscopy has high diagnostic accuracy for upper gastrointestinal precancerous lesions, under the largely retrospective, curated conditions that predominate in the current evidence base, with pooled sensitivity of 0.930, specificity of 0.944, and AUC of 0.9661. Among the variables formally evaluated by meta-regression, imaging modality was associated with heterogeneity, while diagnostic performance also varied meaningfully by target lesion category, with image-enhanced endoscopy showing significantly higher sensitivity than white-light endoscopy. These findings most directly support a role for AI as a detection-assistance tool during routine screening endoscopy, flagging subtle or easily overlooked mucosal abnormalities for endoscopist review, rather than as a stand-alone diagnostic or treatment-planning tool, particularly given the predominance of retrospective, single-center studies with non-consecutive image selection and the current scarcity of real-time, multicenter prospective validation. Future research should prioritize multicenter prospective trials with consecutive patient enrollment, standardized outcome reporting encompassing real-time video evaluation and lesion-subtype-specific performance, and the development of interpretable AI systems to facilitate a more confident transition from experimental research to routine clinical practice in appropriate use scenarios.

Supporting information

Supplementary material for this article is available at https://doi.org/10.14218/CSP.2026.00010.

Acknowledgments

None.

Funding

This work was supported by the Noncommunicable Chronic Diseases–National Science and Technology Major Project (Grant Nos. 2025ZD0545303, 2025ZD0545300), the CAMS Innovation Fund for Medical Sciences (CIFMS) (Grant No. 2025-I2M-XH-XX-047), and the National High-Level Hospital Clinical Research Funding Program (Grant No. 2025-LYZX-R-B04). The funders had no role in the study design; data collection, analysis, or interpretation; manuscript preparation; or the decision to submit the manuscript for publication.

Conflict of interest

The authors declare that they have no conflicts of interest.

Author contributions

Conceptualization (YB, TG, JL), methodology (YB, DH, TG), software (YB, DH), validation (JyL), formal analysis (YB, DH), investigation (YB, DH), data curation (YB, DH, JyL), writing – original draft (YB, DH, YL), writing – review and editing (JyL, YL, JZ, TG, JL), visualization (YL), supervision (JZ, TG, JL), project administration (JZ, JL), and funding acquisition (JL). All authors have made significant contributions to this study and have approved the final manuscript.

Data sharing statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, *et al*. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229–263. DOI: 10.3322/caac.21834, PMID: 38572751.

[2] Arnold M, Abnet CC, Neale RE, Vignat J, Giovannucci EL, McGlynn KA, *et al*. Global Burden of 5 Major Types of Gastrointestinal Cancer. Gastroenterology 2020;159(1):335–349.e15. DOI: 10.1053/j.gastro.2020.02.068, PMID: 32247694.

[3] Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Nikšić M, *et al*. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet 2018;391(10125):1023–1075. DOI: 10.1016/S0140-6736(17)33326-3, PMID: 29395269.

[4] Katai H, Ishikawa T, Akazawa K, Isobe Y, Miyashiro I, Oda I, *et al*. Five-year survival analysis of surgically resected gastric cancer cases in Japan: a retrospective analysis of more than 100,000 patients from the nationwide registry of the Japanese *Gastric Cancer* Association (2001-2007). Gastric Cancer 2018;21(1):144–154. DOI: 10.1007/s10120-017-0716-7, PMID: 28417260.

[5] Correa P. Human gastric carcinogenesis: a multistep and multifactorial process--First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention. Cancer Res 1992;52(24):6735–6740. PMID: 1458460.

[6] de Vries AC, van Grieken NC, Looman CW, Casparie MK, de Vries E, Meijer GA, *et al*. Gastric cancer risk in patients with premalignant gastric lesions: a nationwide cohort study in the Netherlands. Gastroenterology 2008;134(4):945–952. DOI: 10.1053/j.gastro.2008.01.071, PMID: 18395075.

[7] Shaheen NJ, Richter JE. Barrett's oesophagus. Lancet 2009;373(9666):850–861. DOI: 10.1016/S0140-6736(09)60487-6, PMID: 19269522.

[8] Li H, Zhang S, Zhou J, Tong F, Gong J, Zha Z, *et al*. Endoscopic Surveillance for Premalignant Esophageal Lesions: A Community-Based Multicenter, Prospective Cohort Study. Clin Gastroenterol Hepatol 2023;21(3):653–662.e8. DOI: 10.1016/j.cgh.2022.04.039, PMID: 35623589.

[9] Pimentel-Nunes P, Libânio D, Marcos-Pinto R, Areia M, Leja M, Esposito G, *et al*. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal *Endoscopy* (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy 2019;51(4):365–388. DOI: 10.1055/a-0859-1883, PMID: 30841008.

[10] Sharma P, Dent J, Armstrong D, Bergman JJ, Gossner L, Hoshihara Y, *et al*. The development and validation of an endoscopic grading system for Barrett's esophagus: the Prague C & M criteria. Gastroenterology 2006;131(5):1392–1399. DOI: 10.1053/j.gastro.2006.08.032, PMID: 17101315.

[11] Muto M, Minashi K, Yano T, Saito Y, Oda I, Nonaka S, *et al*. Early detection of superficial squamous cell carcinoma in the head and neck region and esophagus by narrow band imaging: a multicenter randomized controlled trial. J Clin Oncol 2010;28(9):1566–1572. DOI: 10.1200/JCO.2009.25.4680, PMID: 20177025.

[12] Guo L, Xiao X, Wu C, Zeng X, Zhang Y, Du J, *et al*. Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). Gastrointest Endosc 2020;91(1):41–51. DOI: 10.1016/j.gie.2019.08.018, PMID: 31445040.

[13] Dilaghi E, Lahner E, Annibale B, Esposito G. Systematic review and meta-analysis: Artificial intelligence for the diagnosis of gastric precancerous lesions and Helicobacter pylori infection. Dig Liver Dis 2022;54(12):1630–1638. DOI: 10.1016/j.dld.2022.03.007, PMID: 35382973.

[14] Hashimoto R, Requa J, Dao T, Ninh A, Tran E, Mai D, *et al*. Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in Barrett's esophagus (with video). Gastrointest Endosc 2020;91(6):1264–1271.e1. DOI: 10.1016/j.gie.2019.12.049, PMID: 31930967.

[15] Wu L, Wang J, He X, Zhu Y, Jiang X, Chen Y, *et al*. Deep learning system compared with expert endoscopists in predicting early gastric cancer and its invasion depth and differentiation status (with videos). Gastrointest Endosc 2022;95(1):92–104.e3. DOI: 10.1016/j.gie.2021.06.033, PMID: 34245752.

[16] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. DOI: 10.1136/bmj.n71, PMID: 33782057.

[17] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529–536. DOI: 10.7326/0003-4819-155-8-201110180-00009, PMID: 22007046.

[18] Plana MN, Arevalo-Rodriguez I, Fernández-García S, Soto J, Fabregate M, Pérez T, *et al*. Meta-DiSc 2.0: a web application for meta-analysis of diagnostic test accuracy data. BMC Med Res Methodol 2022;22(1):306. DOI: 10.1186/s12874-022-01788-2, PMID: 36443653.

[19] Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. BMC Med Res Methodol 2006;6:31. DOI: 10.1186/1471-2288-6-31, PMID: 16836745.

[20] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7(3):177–188. DOI: 10.1016/0197-2456(86)90046-2, PMID: 3802833.

[21] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med 2004;23(9):1351–1375. DOI: 10.1002/sim.1761, PMID: 15116347.

[22] Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58(10):982–990. DOI: 10.1016/j.jclinepi.2005.02.022, PMID: 16168343.

[23] Viechtbauer W, Cheung MW. Outlier and influence diagnostics for meta-analysis. Res Synth Methods 2010;1(2):112–125. DOI: 10.1002/jrsm.11, PMID: 26061377.

[24] Lin L, Chu H. Quantifying publication bias in meta-analysis. Biometrics 2018;74(3):785–794. DOI: 10.1111/biom.12817, PMID: 29141096.

[25] Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Syst Rev 2022;18(2):e1230. DOI: 10.1002/cl2.1230, PMID: 36911350.

[26] Liu G, Hua J, Wu Z, Meng T, Sun M, Huang P, *et al*. Automatic classification of esophageal lesions in endoscopic images using a convolutional neural network. Ann Transl Med 2020;8(7):486. DOI: 10.21037/atm.2020.03.24, PMID: 32395530.

[27] Wong PK, Yao L, Yan T, Choi IC, Yu HH, Hu Y. Broad learning system stacking with multi-scale attention for the diagnosis of gastric intestinal metaplasia. Biomed Signal Process Control 2022;73:103476. DOI: 10.1016/j.bspc.2021.103476.

[28] Gong EJ, Bang CS, Jung K, Kim SJ, Kim JW, Seo SI, *et al*. Deep-Learning for the Diagnosis of Esophageal Cancers and Precursor Lesions in Endoscopic Images: A Model Establishment and Nationwide Multicenter Performance Verification Study. J Pers Med 2022;12(7):1052. DOI: 10.3390/jpm12071052, PMID: 35887549.

[29] Mu G, Zhu Y, Niu Z, Li H, Wu L, Wang J, *et al*. Expert-level classification of gastritis by endoscopy using deep learning: a multicenter diagnostic trial. Endosc Int Open 2021;9(6):E955–E964. DOI: 10.1055/a-1372-2789, PMID: 34079883.

[30] Liu X, Wang C, Bai J, Liao G. Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images. Neurocomputing 2020;392:253–267. DOI: 10.1016/j.neucom.2018.10.100.

[31] Jhang JY, Tsai YC, Hsu TC, Huang CR, Cheng HC, Sheu BS. Gastric Section Correlation Network for Gastric Precancerous Lesion Diagnosis. IEEE Open J Eng Med Biol 2024;5:434–442. DOI: 10.1109/OJEMB.2023.3277219, PMID: 38899022.

[32] Tan JL, Pitawela D, Chinnaratha MA, Beany A, Aguila EJ, Chen HT, *et al*. Exploring vision transformers for classifying early Barrett's dysplasia in endoscopic images: A pilot study on white-light and narrow-band imaging. JGH Open 2024;8(9):e70030. DOI: 10.1002/jgh3.70030, PMID: 39328302.

[33] Siripoppohn V, Pittayanon R, Tiankanon K, Faknak N, Sanpavat A, Klaikaew N, *et al*. Real-time semantic segmentation of gastric intestinal metaplasia using a deep learning approach. Clin Endosc 2022;55(3):390–400. DOI: 10.5946/ce.2022.005, PMID: 35534933.

[34] Li Z, Zheng X, Mu Y, Zhang M, Liu G. The intelligent gastrointestinal metaplasia assessment based on deformable transformer with token merging. Biomed Signal Process Control 2024;95(Pt B):106454. DOI: 10.1016/j.bspc.2024.106454.

[35] Chiang TH, Hsu YN, Chen MH, Chen YR, Cheng HC, Chen MJ, *et al*. A rural-to-center artificial intelligence model for diagnosing

Helicobacter pylori infection and premalignant gastric conditions using endoscopy images captured in routine practice. Endoscopy 2026;58(4):343–354. DOI: 10.1055/a-2721-6552, PMID: 41082919.

[36] Gong EJ, Bang CS, Lee JJ. Edge Artificial Intelligence Device in Real-Time Endoscopy for Classification of Gastric Neoplasms: Development and Validation Study. Biomimetics (Basel) 2024;9(12): 783. DOI: 10.3390/biomimetics9120783, PMID: 39727787.

[37] Lin N, Yu T, Zheng W, Hu H, Xiang L, Ye G, *et al*. Simultaneous Recognition of Atrophic Gastritis and Intestinal Metaplasia on White Light Endoscopic Images Based on Convolutional Neural Networks: A Multicenter Study. Clin Transl Gastroenterol 2021;12(8):e00385. DOI: 10.14309/ctg.0000000000000385, PMID: 34342293.

[38] Yang J, Ou Y, Chen Z, Liao J, Sun W, Luo Y, *et al*. A Benchmark Dataset of Endoscopic Images and Novel Deep Learning Method to Detect Intestinal Metaplasia and Gastritis Atrophy. IEEE J Biomed Health Inform 2023;27(1):7–16. DOI: 10.1109/JBHI.2022.3217944, PMID: 36306301.

[39] Xu M, Zhou W, Wu L, Zhang J, Wang J, Mu G, *et al*. Artificial intelligence in the diagnosis of gastric precancerous conditions by image-enhanced endoscopy: a multicenter, diagnostic study (with video). Gastrointest Endosc 2021;94(3):540–548.e4. DOI: 10.1016/j.gie.2021.03.013, PMID: 33722576.

[40] de Groof AJ, Struyvenberg MR, van der Putten J, van der Sommen F, Fockens KN, Curvers WL, *et al*. Deep-Learning System Detects Neoplasia in Patients With Barrett's Esophagus With Higher Accuracy Than Endoscopists in a Multistep Training and Validation Study With Benchmarking. Gastroenterology 2020;158(4): 915–929.e4. DOI: 10.1053/j.gastro.2019.11.030, PMID: 31759929.

[41] Luo H, Xu G, Li C, He L, Luo L, Wang Z, *et al*. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. Lancet Oncol 2019;20(12):1645–1654. DOI: 10.1016/S1470-2045(19)30637-0, PMID: 31591062.

[42] Yuan XL, Liu W, Lin YX, Deng QY, Gao YP, Wan L, *et al*. Effect of an artificial intelligence-assisted system on endoscopic diagnosis of superficial oesophageal squamous cell carcinoma and precancerous lesions: a multicentre, tandem, double-blind, randomised controlled trial. Lancet Gastroenterol Hepatol 2024;9(1):34–44. DOI: 10.1016/S2468-1253(23)00276-5, PMID: 37952555.

[43] Renna F, Martins M, Neto A, Cunha A, Libânio D, Dinis-Ribeiro M, *et al*. Artificial Intelligence for Upper Gastrointestinal Endoscopy: A Roadmap from Technology Development to Clinical Practice. Diagnostics (Basel) 2022;12(5):1278. DOI: 10.3390/diagnostics12051278, PMID: 35626433.

[44] Meinikheim M, Mendel R, Palm C, Probst A, Muzalyova A, Scheppach MW, *et al*. Influence of artificial intelligence on the diagnostic performance of endoscopists in the assessment of Barrett's esophagus: a tandem randomized and video trial. Endoscopy 2024;56(9):641–649. DOI: 10.1055/a-2296-5696, PMID: 38547927.

[45] Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, *et al*. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020;10(1): 12598. DOI: 10.1038/s41598-020-69250-1, PMID: 32724046.