Surgical Endoscopy (2026) 40:1651–1660
https://doi.org/10.1007/s00464-025-12410-z

Artificial intelligence assistance narrows the experience gap in endoscopic reporting of gastric lesions: a prospective clinical trial

Hang You[1,2,3,4] · Xiao Tao[1,2,3,4] · Zehua Dong[1,2,3,4] · Yanxia Li[1,2,3,4] · Li Huang[1,2,3,4] · Jia Li[1,2,3,4] · Hongliu Du[1,2,3,4] · Mei Deng[1,2,3,4] · Zhifeng Wu[1,2,3,4] · Xia Tan[1,2,3,4] · Ting Yang,[1,2,3,4] · Jun Liu[1,2,3,4] · Honggang Yu[1,2,3,4]

Received: 29 June 2025 / Accepted: 11 November 2025 / Published online: 4 December 2025
© The Author(s) 2025

Abstract

Background High-quality endoscopic reports are crucial for clinical decision-making in gastric lesions. However, manually generated reports often suffer from inconsistencies, resulting in incomplete lesion documentation and insufficient descriptions. The objective is to validate an artificial intelligence (AI) reporting system for improving the quality of gastric lesion documentation with a particular focus on narrowing the experience gap among endoscopists.

Methods We retrospectively collected gastric images to develop a reporting system for gastric lesions. The system comprises several deep learning models for lesion detection, classification, and feature recognition, and was validated using 276 video clips. Both retrospective and prospective case validations were conducted to evaluate the system's clinical effectiveness, comparing the completeness of reports with and without AI assistance, particularly in describing features of suspicious lesions.

Results In video validation, the system identified 99.27% (274/276) of lesions, with 88.97% accuracy in neoplasm detection, achieving 93.75% sensitivity and 86.98% specificity. Retrospective analyses showed that junior endoscopists using AI reported significantly more complete lesion documentation than original reports (74.75% *vs* 58.08%, $P<0.001$). In the prospective trial, the AI-assisted group demonstrated superior lesion reporting completeness (86.50% *vs* 75.11%, $P=0.001$). Notably, the AI group reported all 18 identified suspicious neoplasms, while the routine group reported only 16.

Conclusions This study confirms the efficacy of an AI reporting system in documenting gastric focal lesions, highlighting significant improvements in the completeness and accuracy of reports generated by endoscopists.

Keywords Gastric lesions · Artificial intelligence · Endoscopy reporting

Gastroscopy serves as an essential tool for the detection and diagnosis of gastric lesions [1, 2]. An endoscopic diagnostic report summarizes and condenses the findings of a gastroscopic examination. A comprehensive and accurate report typically includes a diagnosis and a detailed description of observed lesions, forming the cornerstone for subsequent therapeutic decisions, follow-up strategies, and effective clinician communication [3].

Specific organizations, such as the European Society of Gastrointestinal Endoscopy, have established requirements for the information that should be included in reports. However, the quality of manually generated endoscopy reports varies. In addition to incomplete lesion records, the descriptions of lesion characteristics are often overlooked [1, 4]. In a retrospective study, among 33 patients documented with esophagitis, only 57% of the reports recorded the length of columnar-lined epithelium, and none of the reports described the Prague classification system [5].

Hang You, Xiao Tao, Zehua Dong these authors are contributed equally to this work.

✉ Jun Liu
2451938436@qq.com

✉ Honggang Yu
yuhonggang@whu.edu.cn

[1] Department of Gastroenterology, Renmin Hospital of Wuhan University, 99 Zhangzhidong Road, Wuhan 430060, Hubei Province China, China

[2] Hubei Provincial Clinical Research Center for Digestive Disease Minimally Invasive Incision, Renmin Hospital of Wuhan University, Wuhan, China

[3] Key Laboratory of Hubei Province for Digestive System Disease, Renmin Hospital of Wuhan University, Wuhan, China

[4] Engineering Research Center for Artificial Intelligence Endoscopy Interventional Treatment of Hubei Province, Wuhan, China

Despite the clinical importance of standardized endoscopic reporting guidelines, their practical implementation remains inconsistent. With the advancement of artificial intelligence (AI) technology, it has become feasible to employ AI for meticulous recording of procedure findings, thereby enhancing the comprehensiveness and accuracy of reports. In our previous study, we developed an AI-based endoscopy automatic reporting system (AI-EARS) that effectively generates standardized photographic and textual reports during endoscopic examinations [6, 7]. However, the study did not further identify or describe the features of the gastric lesions, and there is still a lack of evidence for applying such a system in a real clinical setting. Therefore, this study aimed, for the first time, to further validate the effectiveness of an AI-assisted system in enhancing the documentation of focal gastric lesions through a prospective back-to-back clinical trial, in order to explore the feasibility of implementing such an intelligent reporting system in real-world clinical practice.

Materials and methods

Datasets

Dataset 1: We included 276 video clips at Renmin Hospital of Wuhan University for lesion diagnosis and feature recognition. The video clips included xanthelasma ($n=31$), submucosal tumors (SMT $n=15$), polyps ($n=59$), low-risk ulcers ($n=13$), erosions, and gastric neoplasms ($n=78$). Lesions were selected based on the following exclusion criteria: the presence of multiple lesions within the same field of view; poor field of view of the image.

Dataset 2: We retrospectively included 97 consecutive patients from Renmin Hospital of Wuhan University between May 09, 2022, and May 23, 2023. Their complete gastroscopy videos and original reports were collected to explore the applicability of the AI system.

All endoscopic images used for training and validation were obtained in white light view from multiple medical centers. Instruments used in this study included gastroscopes from two vendors (Olympus Optical Co., Tokyo, Japan; Fujifilm, Co., Kanagawa, Japan). The datasets, training, and testing results were Supplementary Materials.

Development of the models and system construction

The real-time gastroscopy reporting system consists of four integrated models. Model 1 detects focal gastric lesions and ensures standard photo-documentation for each lesion. Model 2 identifies lesions as either neoplasm or non-neoplasm, marked with a red or blue box, respectively. Model 3 categorizes the type of lesion, while Model 4 identifies and outputs the specific features of suspicious neoplasms. The gastroscopy video signal is decoded at 25 fps for real-time analysis and prediction, and a structured text report is generated immediately upon completion of the procedure.The input image resolution was 416×416 pixels. All artificial intelligence models were trained using open-source deep learning frameworks (TensorFlow 1.12.2, Keras 2.1.5, and PyTorch 2.0.1), with high-performance computing support from an NVIDIA GeForce RTX 3080 Ti GPU. The inference time per frame was within 40 ms, and approximately 15% of frames were considered unevaluable due to poor visualization (e.g., obscured by bleeding or poor views).

Model 1 was trained and tested using YOLO v3, while Model 2 employed ResNet-50. The detailed construction and performance of these models have been described in previous studies. Model 3 was also developed based on ResNet-50 to classify low-risk gastric lesions. The classification accuracies for gastric ulcer, xanthelasma, erosion, benign polyps, and submucosal tumors were 87.41, 99.32, 95.97, 95.42, and 95.54%, respectively, as previously reported. Model 4 was trained and tested on a dedicated dataset, to identify six types of focal lesion features (including location, bleeding, boundary, mucosal roughness, color, and morphology).

Video validation of the system

Dataset 1 was utilized for video validation to explore the system's comprehensive performance in lesion detection, classification, and feature characterization. First, we evaluated the system's accuracy in detecting lesions. Next, we assessed its performance in classifying lesions as neoplasms or non-neoplasms. Finally, for suspicious focal lesions such as neoplasms and erosions, we evaluated its ability to describe lesion features.

Retrospective case validation

Five junior endoscopists (with less than five years of gastroscopy experience), who were not involved in the construction of the AI-EARS, data collection, or dataset preprocessing, independently reviewed gastroscopy procedures from Dataset 2. They were required to review the entire gastroscopy videos and draft reports with assistance from the AI system. We then compared the completeness of lesion labeling, the comprehensiveness of feature descriptions, and the accuracy of these descriptions between reports generated by junior endoscopists using the AI system and the original reports generated by senior endoscopists.

Prospective back-to-back clinical trial

We prospectively recruited patients undergoing gastroendoscopy at the Renmin Hospital of Wuhan University using the following inclusion criteria: (1) age 18 years or older, (2) sedated upper endoscopy, and (3) be able to read, understand, and sign an informed consent. The exclusion criteria were as follows: (1) absolute contraindications to EGD examination; (2) not suitable for sedated endoscopy assessed by anesthesiologists; (3) contraindications to biopsy; (4) known active upper gastrointestinal bleeding or emergency endoscopy; (5) pregnancy; (6) previous gastric surgery; (7) food residue; and (8) researchers believe that the patient is not suitable to participate in the trial (e.g., severe comorbidities, impaired compliance, or psychiatric/cognitive disorders). Nine endoscopists from Renmin Hospital of Wuhan University, including four senior (with EGD experience 5–10 years) and five junior endoscopists (with EGD experience < 5 years), participated in this prospective observational study. Junior endoscopists generated reports using the AI-assisted system, while senior endoscopists followed the conventional workflow. In the conventional workflow, reports were generated using a fixed text template structured by gastric regions (fundus, body, angle, and antrum). Each region served as a predefined header with corresponding text fields. Endoscopists filled in or modified these fields based on the patient's findings, rather than composing entirely de novo free-text reports. Both groups independently observed the same endoscopic procedures and drafted their reports in a back-to-back manner. They were blinded to each other's findings, and no verbal communication was allowed throughout the process, ensuring a fair and unbiased comparison under identical clinical conditions. When the reports were completed, they were saved on a computer and collected by a research assistant. Additionally, the original video of each case was preserved for expert review to establish the gold standard.

The primary endpoints encompassed the completeness of lesions reporting. The secondary outcomes included the completeness of report drafting on lesion features, the accuracy of report drafting on lesion features, reporting time, completeness of reporting lesions using an AI system, and the accuracy of report drafting on lesion features of an AI system. This report followed the guidelines of the CONSORT-AI statement.

Gold standard

For lesions in retrospective videos, case validation, and prospective clinical trials, the gold standard for diagnosis is pathological diagnosis for neoplasms. Non-neoplastic lesions and areas without lesions were classified by three expert endoscopists, with discrepancies resolved by discussion and voting. For feature recognition, the gold standard was defined by the same panel using this procedure.

For cases in the retrospective and prospective cohort, the three experts were invited to review all procedure videos by frame and decide the lesions supposed to be reported, and suspicious neoplasms of which the features were supposed to be identified after the videos had been collected. The experts stated and discussed each opinion, then voted on the different decisions until they reached a consensus. The majority decision was regarded as the final consensus when disagreement persisted.

Ethics

The ethics of the prospective trial were approved by the Ethics Committee of Renmin Hospital of Wuhan University (approval No.WDRY2019-K073X03). The study was registered on ClinicalTrails.gov (Identifier: NCT05851716; registration date: April 30, 2023; full title: "A Prospective, Single-arm, Single-center Observational Study to Evaluate the Clinical Effectiveness of an Intelligent Reporting System for the Upper Gastrointestinal Endoscopy"; URL: https://clinicaltrials.gov/study/NCT05851716

Statistical analysis

The baseline characteristics of the cohort are presented as mean ± SD or number (percentage). The completeness of the lesion capture evaluated the system's performance on the endoscopy report, the accuracy of the location of the lesion, the accuracy of the lesion type, and the accuracy of the report (including both lesions and lesion types). Completeness of reporting was the percentage of accurately reported lesions relative to the total number of lesions. Similarly, completeness of report drafting concerning lesion features was defined as the percentage of drafted features compared to the total number of features required for accurate reporting. Furthermore, the accuracy of lesion feature reporting was defined as the percentage of features accurately drafted against the total number of drafted features. McNemar's statistic was used to evaluate the completeness of abnormality reports for each endoscopist. Chi-square tests and Fisher's exact test were used to assess the accuracy of lesion characteristics in the reports. A significance level of 0.05 was used to compare the endoscopists' performance with and without AI assistance.

Results

Performance of the system for gastric lesion classification and feature recognition in video clips

In video validation, the system automatically captured 99.27% (274/276) of the gastric focal lesions. In identifying neoplasms, the accuracy was 88.97% (242/272), with a sensitivity and specificity of 93.75% (75/80) and 86.98% (167/192), respectively.

Among the 80 gastric neoplasms and 78 erosions observed in this dataset, AI achieved an accuracy of 85.44% (135/158) for bleeding, 81.01% (128/158) for boundary, 79.11% (125/158) for mucosal roughness, 68.99% (109/158) for color, and 63.29% (100/158) for morphology recognition (Table 1). The definitions and categorical values of each feature are provided in Supplementary Table 2.

Reporting under AI assistance in the retrospective cohort

Under AI assistance, the completeness of lesion reporting of junior endoscopists was significantly higher than that of original reports (74.75% *vs* 58.08%, $P<0.001$). Table 2 and Fig. 2 show the completeness for each focal lesion. Regarding erosions, the original reports only reported 53.06% (52/98), significantly lower than the 87.76% (86/98) reported by junior endoscopists ($P<0.001$).

With regard to describing features of the lesions, the completeness of report drafting on lesion features by junior endoscopists was 98.85%, higher than the original report of 33.33% ($P<0.001$). Compared with the original reports, the completeness of boundary (100.00% *vs* 14.29%, $P<0.001$) and color (100.00% *vs* 42.86%, $P=0.005$) was significantly improved. When analyzing the accuracy of report drafting on lesion features, junior endoscopists achieved 98.85% (39/42), and the original reports achieved 85.71% (12/14).

Reporting under AI assistance in the prospective clinical trial

A total of 108 patients undergoing gastroendoscopy from September 04, 2023, to September 19, 2023, were recruited to assess the applicability of the AI system. Among these participants, 8 patients were excluded as they did not fulfill the inclusion and exclusion criteria (Fig. 1). Therefore, 100 patients were included in the analysis. Baseline information of these patients is shown in Table 3.

Comparison of endoscopists' performance

The overall completeness of lesion reporting in the AI-assisted group was significantly higher than that in the routine group (86.50% *vs* 75.11%, $P=0.001$). Particularly in reporting focal erosion, the AI-assisted group was 85.42% more complete than the 72.73% in the routine group ($P=0.037$) (Table 4 and Fig. 2).

A total of 18 suspicious neoplasms were identified, with their features confirmed by experts, of which 14 underwent biopsy, and detailed histological diagnoses are provided in the Supplementary Table 1. The AI-assisted group successfully reported all 18 lesions, whereas the routine group reported only 16. The overall completeness of lesion feature reporting in the

Table 1 Video validation results of the endoscopic automatic image reporting system

	Neoplasm	Erosion	Xanthoma	SMT	Polyp	Ulcer	Total
Detection rate	100.00%;80/80 (95.42–100)	100%;78/78 (95.31–100)	96.77%;30/31 (83.80–99.43)	100%;13/13 (77.19–100)	98.31%;58/59 (91.01–99.7)	100.00%;13/13 (77.19–100)	99.27%;274/276 (97.40–99.80)
Accuracy	93.75%;75/80 (86.19–97.30)	74.36%;58/78 (63.69–82.75)	96.67%;29/30 (83.33–99.41)	100%;13/13 (77.19–100)	98.28%;57/58 (90.87–99.70)	76.92%;10/13 (49.74–91.82)	88.97%;242/272 (84.69–92.16)
Location	66.25%;53/80 (55.36–75.65)	82.05%;64/78 (72.10–89.00)	86.67%;26/30 (70.32–94.69)	69.23%;9/13 (42.37–87.32)	77.19%;44/57 (64.79–86.84)	80.00%;8/10 (49.02–94.33)	76.12%;204/268 (70.67–81.00)
Bleeding	76.25%;61/80 (65.86–84.24)	94.87%;74/78 (87.54–97.99)	–	–	–	–	–
Boundary	91.25%;73/80 (83.02–95.70)	70.51%;55/78 (59.62–79.47)	–	–	–	–	–
Mucosal roughness	86.25%;69/80 (77.03–92.15)	71.79%;56/78 (60.96–80.57)	–	–	–	–	–
Color	65.00%;52/80 (54.08–74.55)	73.08%;57/78 (62.33–81.67)	–	–	–	–	–
Morphology	52.50%;42/80 (41.70–63.08)	74.36%;58/78 (63.69–82.75)	–	–	–	–	–

Table 2 Comparison of completeness of lesion reports with and without AI assistance

		Ulcer	SMT	Erosion	Xanthoma	Polyp	Total
Retrospective	AI-assisted	100.00%;4/4 (51.01–100)	66.67%;4/6 (30.00–90.32)	87.76%;86/98 (79.87–92.98)	42.86%;3/7 (15.82–74.95)	62.65%;52/83 (51.76–72.56)	74.75%;148/198 (68.31–80.39)
	Routine	75.00%;3/4 (30.06–95.44)	83.33%;5/6 (43.65–97.04)	53.06%;52/98 (43.18–62.72)	28.57%;2/7 (8.22–64.12)	59.04%;49/83 (48.19–69.22)	58.08%;115/198 (51.17–64.73)
	P	-	1	< 0.001	1	0.585	< 0.001
Prospective	AI	100.00%;2/2 (34.24–100)	66.67%;6/9 (35.88–87.88)	80.42%;115/143 (73.20–86.24)	55.56%;5/9 (26.67–81.10)	86.49%;64/74 (76.92–92.64)	80.59%;192/237 (75.18–85.20)
	AI-assisted	100.00%;2/2 (34.24–100)	77.78%;7/9 (45.30–93.68)	85.42%;123/143 (78.73–90.44)	55.56%;5/9 (26.67–81.10)	91.89%;68/74 (83.41–96.26)	86.50%;205/237 (81.71–90.25)
	Routine	100.00%;2/2 (34.24–100)	88.89;8/9 (56.50–98.02)	72.73%;104/143 (64.90–79.49)	66.67%;6/9 (35.88–87.88)	78.38%;58/74 (67.65–86.33)	75.11%;178/237 (69.35–80.17)
	*P**	–	1	0	0.500	0	0
	*P***	–	0.781	0.708	0.938	0.018	0.056
	*P****	–	1	0.069	0.162	0.056	0.004

Data are as %; n/N (95% CI)

P^{*} indicates the comparison between AI and AI-assisted group

P^{**} indicates the comparison between AI and routine group

P^{***} indicates the comparison between AI-assisted group and routine group

AI-assisted group was 50.00% (54/108), significantly higher than 21.88% (21/96) in the routine group ($P < 0.001$). Specifically, the AI-assisted group showed significant improvements in reporting completeness for mucosal roughness (55.56% *vs* 18.75%, $P = 0.028$) compared to the routine group.

The accuracy of report drafting on lesion features in the AI-assisted and routine group was 72.97% (54/74) and 60.00% (21/35), respectively, with no statistically significant difference between them (Table 5). The average time taken by junior endoscopists to draft a report with the AI reporting system's assistance was 212.10 ± 119.16 s (median 179.0 s; IQR 167.5 s, range 14–623 s), compared to 203.62 ± 71.70 s (median 190.5 s; IQR 101.25 s, range 77–470 s) for senior endoscopists using a traditional reporting system, with no statistically significant difference ($P = 0.425$).

In addition, we also evaluated the independent performance of the AI system in reporting focal lesions and describing the features of lesions. Among 237 focal lesions, the AI system successfully reported 80.59% (192/237) of these lesions, which was higher than the 75.11% (178/237) reported by senior endoscopists, but lower than the 86.50% (205/237) reported by junior endoscopists with AI assistance ($P < 0.001$).

Discussion

This study is the first to evaluate an automated AI-based endoscopy reporting system using a back-to-back trial design and to validate its performance in real-world clinical settings. An objective comparison between the AI-assisted and conventional reporting methods was achieved through a controlled study involving independently operating junior and senior endoscopists under standardized clinical conditions. The results demonstrated that the AI system significantly enhanced the reporting quality of junior endoscopists, elevating it to a level comparable to or exceeding that of senior endoscopists. These empirical findings not only confirm the clinical applicability of the AI system but also provide a robust foundation for the broader adoption of AI technologies in the standardization and quality improvement of medical documentation.

Manually generating endoscopy reports is time-consuming and labor-intensive, leading to inconsistent report quality. This variability hinders effective information exchange and high-quality interdisciplinary care. Consequently, the promotion of standardized endoscopic examination reporting has been a longstanding objective. In this context, an automated endoscopy report generation system presents an ideal solution (3). In a previous study, we developed an intelligent system for automatic photo-documentation and tested its utility at the image level. In a subsequent study, we expanded the system's functionality by incorporating the generation of structured text reports, thereby creating a comprehensive automated reporting system. The system's effectiveness was validated using a multi-reader, multi-case approach [6, 7]. However, the applicability of such an AI reporting system in a real clinical setting has not yet been evaluated. In this study, we incorporated the functionality to identify and describe the features of suspicious neoplasms and validated this capability using both images and videos. Additionally, we performed a retrospective cohort study and a prospective clinical trial to evaluate the drafting of reports for gastric focal lesions. These investigations provided insights into the potential role of an AI reporting system

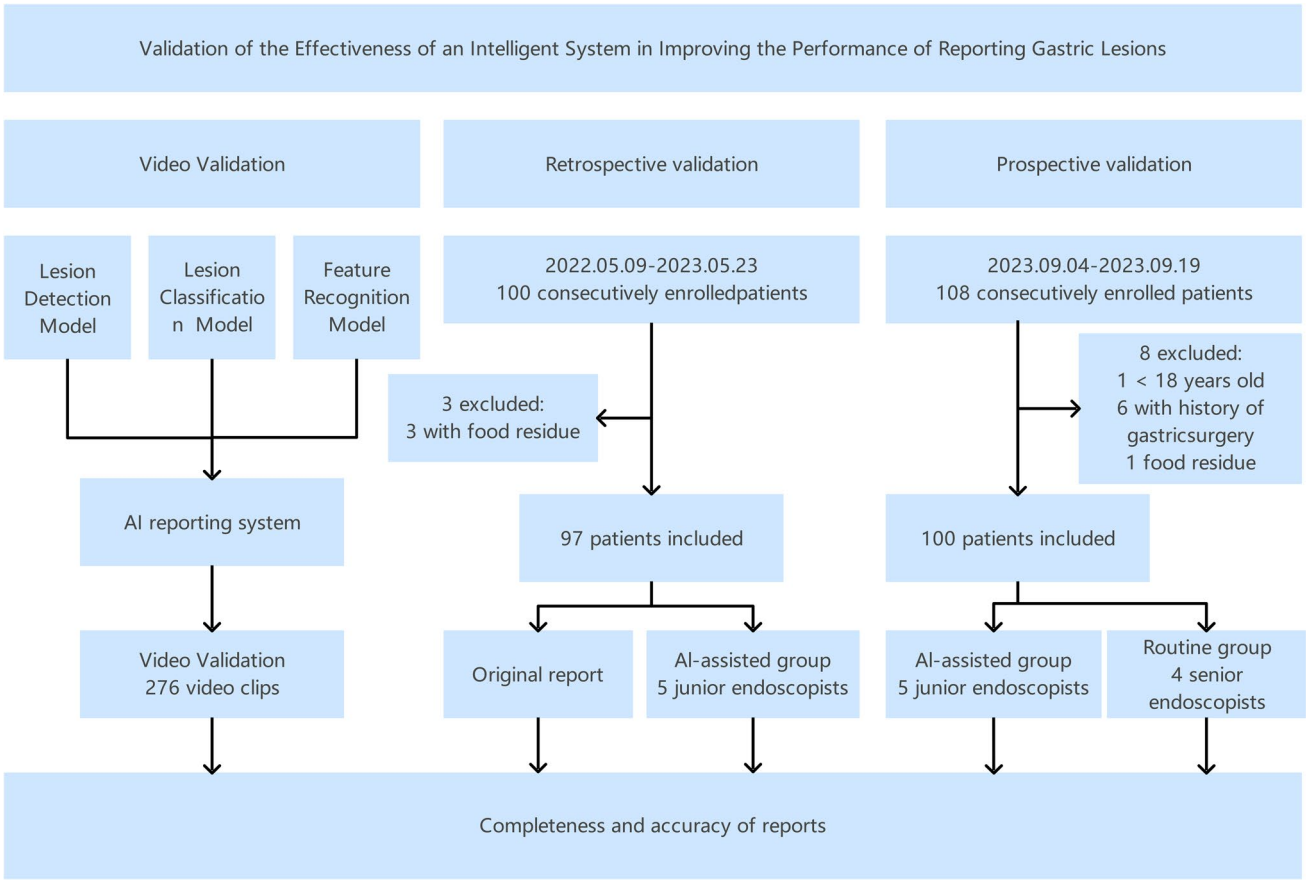


Fig. 1 Workflow diagram for the video, retrospective, and prospective validation

Table 3 Baseline characteristics of patients in the retrospective and prospective validation

	Retrospective validation	Prospective validation
Age, *n* (±SD)	49.392 ± 13.924	52.01 ± 15.155
Sex		
Male	48 (49.5)	36 (36.0)
Female	49 (50.5)	64 (64.0)
Neoplasm	0	0
Advanced cancer	0	0
EC	0	0
Low-grade	0	0
Non-neoplasm	198	244
Ulcer, *n*, (%)	4 (2.02)	2 (0.8)
SMT, *n*, (%)	6 (3.03)	9 (3.8)
Erosion, *n*, (%)	98 (49.49)	143 (58.3)
Xanthoma, *n*, (%)	7 (3.54)	9 (3.8)
Polyp, *n*, (%)	83 (41.92)	74 (30.2)
Vasodilation, *n*, (%)	0 (0)	7 (2.9)
Total	198	244

in aiding junior endoscopists with report drafting, thereby enhancing the quality of gastroscopy reports.

Gastroscopic procedures necessitate detailed visualization of the gastric mucosa, with a focus on detecting and

identifying lesions. For suspected neoplasms, a detailed description is crucial not only for accurate diagnosis but also for informing subsequent therapeutic decisions [8–12]. A comprehensive and accurate report of common focal lesions demonstrates the diagnostic skills and professionalism of the endoscopist. In our prospective validation, all 14 biopsied suspicious lesions were confirmed as non-neoplastic, which was expected given the low prevalence of true neoplasia in a consecutively enrolled cohort of 100 patients. In clinical practice, many non-neoplastic lesions exhibit neoplasm-like features [13–15]; false positives are therefore not only an inherent challenge but also unavoidable when aiming to minimize the risk of missed early neoplasia [16]. In an earlier study, we developed an explainable AI system that incorporates domain knowledge through feature extraction and multi-feature fitting methods [17]. This system demonstrated greater clinical credibility and acceptance compared to traditional deep learning approaches and significantly enhanced the diagnostic capabilities of endoscopists. Therefore, this evidence shows that providing a detailed description of the features of suspected neoplasms is essential. Standardized lesion descriptions enable physicians to accurately locate and assess lesions, thereby minimizing omissions and establishing a robust foundation for qualitative endoscopic diagnosis [18]. This approach not only enhances the completeness of patient reports and clarifies the patient's condition but also fosters effective interdisciplinary communication.

Lack of training in gastroscopy procedures has led to significant differences in lesion diagnosis between junior and senior endoscopists, as observed in various diagnostic tasks. Numerous studies have demonstrated that with the assistance of AI systems, junior endoscopists can achieve diagnostic capabilities comparable to those of experts [19–22]. He et al. showed that the accuracy and sensitivity of endoscopists in diagnosing early gastric cancer were significantly improved with the assistance of an AI system [19]. However, there are limited studies exploring the effect of AI systems on assisting junior endoscopists in report drafting. This study addresses this gap by comparing AI-assisted junior endoscopists with senior endoscopists using traditional reporting methods in a prospective clinical validation.

In the retrospective cohort validation, AI-assisted reports drafted by junior endoscopists were significantly more complete compared to the original reports, particularly in identifying erosions and polyps, which were often overlooked in the original reporting documentation. In the prospective trial, senior endoscopists, despite their experience, were prone to missing erosions and polyps in reports. In contrast, the completeness of reports drafted by junior endoscopists significantly improved with the support of the AI reporting system. This indicates that AI can address endoscopists' omissions in lesion reporting through automated analysis

Table 4 Comparison of completeness of lesion features and accuracy of lesion reports with and without AI assistance

		Suspicious neoplasms	Biopsy	Location	Bleeding	Boundary	Mucosal roughness	Color	Morphology	Total
Retrospective	AI-assisted	100.00%;7/7 (64.57–100)	28.57%;2/7 (8.22–64.11)	100.00%;7/7 (64.57–100)	100.00%;7/7 (64.57–100)	100.00%;7/7 (64.57–100)	85.71%;6/7 (48.68–97.43)	100.00%;7/7 (64.57–100)	71.43%;5/7 (35.89–91.78)	98.85%;39/42 (80.99–97.54)
	Routine	100.00%;7/7 (64.57–100)	28.57%;2/7 (8.22–64.11)	85.71%;6/7 (48.68–97.43)	14.29%;1/7 (2.57–51.32)	0;0/7 (0–35.43)	42.86%;3/7 (15.82–74.96)	14.29%;1/7 (2.57–51.32)	42.86%;3/7 (15.82–74.96)	33.33%;14/42 (21.01–48.44)
	P	1	1	1	0.005	0.001	0.265	0.005	0.592	0
Prospective	AI-assisted	100.00%;18/18 (82.41–100)	77.78%;14/18 (54.79–91.0)	66.67%;12/18 (43.75–83.72)	38.89%;7/18 (20.31–61.38)	44.44%;8/18 (24.56–66.28)	55.56%;10/18 (33.72–75.44)	55.56%;10/18 (33.72–75.44)	38.89%;7/18 (20.31–61.38)	50.00%;54/108 (40.73–59.27)
	Routine	88.89%;16/18) (67.2–96.9)	77.78%;14/18 (54.79–91.0)	50.00%;8/16 (28.0–72.0)	6.25%;1/16 (1.11–28.33)	12.5%;2/16 (3.5–36.02)	18.75%;3/16 (6.59–43.01)	25.00%;4/16 (10.18–49.5)	18.75%;3/16 (6.59–43.01)	21.88%;21/96 (14.78–31.14)
	P	0.1456	1	0.324	0.067	0.096	0.028	0.071	0.363	0



Fig. 2 Comparison of report completeness and feature completeness with and without AI assistance

Table 5 Comparison of the accuracy of lesion reports with and without AI assistance

		Location	Bleeding	Boundary	Mucosal roughness	Color	Morphology	Total
Retrospective	AI-assisted	100.00%;7/7 (64.57–100)	100.00%;7/7 (64.57–100)	100.00%;7/7 (64.57–100)	85.71% (6/7)	100.00%;7/7 (64.57–100)	71.43%;5/7 (35.89–91.78)	98.85%;39/42 (80.99–97.54)
	Routine	66.67%;4/6 (30.0–90.32)	100.00%;1/1 (20.65–100)	–	100.00%;3/3 (43.85–100)	100.00%;1/1 (20.65–100)	100.00%;3/3 (43.85–100)	85.71%;12/14 (60.05–95.99)
	P	0.192	–	–	–	–	–	0.787
Prospective	AI-assisted	70.59%;12/17 (46.87–86.72)	87.50%;7/8 (52.91–97.76)	88.89%;8/9 (56.5–98.01)	71.43%;10/14 (45.35–88.28)	76.92%;10/13 (49.74–91.82)	53.85%;7/13 (29.15–76.8)	72.97%;54/74 (61.91–81.77)
	Routine	61.54%;8/13 (35.52–82.29)	100.00%;1/1 (20.65–100)	66.67%;2/3 (20.77–93.85)	42.85%;3/7 (15.82–74.96)	66.67%;4/6 (30.0–90.32)	60.00%;3/5 (23.07–88.24)	60.00%;21/35 (43.57–74.45)
	P	0.896	–	–	0.427	1	–	0.172

and standardized lesion documentation, thereby improving the overall comprehensiveness of reports.

Additionally, we noted that in prospective studies, both the AI-assisted group and the routine group demonstrated higher completeness of reports compared to their performance in retrospective studies (80.59% *vs* 74.75%, 75.11% *vs* 58.08%). In the AI-assisted group, although we prohibited interactions between the AI-assisted group of physicians and other physicians, it was impossible to eliminate discussions among other doctors, which might have led to identifying more lesions. With regard to the routine group, the difference observed between the two rounds of validation may be attributed to the observer effect.

We also observed that the completeness of AI-independent report drafting was better than that of senior endoscopists but lower than that of AI-assisted junior endoscopists. AI in standardized lesion detection and structured data analysis lacks the ability to make clinical contextual judgments and has limitations when used independently. However, when junior doctors combine AI assistance with their clinical experience and patient background, they can correct potential AI errors and improve the completeness of reports. Conversely, senior endoscopists may occasionally overlook details due to their reliance on experience [23]. AI assistance can compensate for this limitation, leading to a more efficient clinical workflow [24, 25]. This highlights that the optimal approach for diagnosis involves synergy between AI and human physicians.

As the first clinical validation of an AI reporting system, there are still several limitations in this study. Firstly, the study was validated only in a single center using endoscopes from specific vendors, which may limit generalizability to other populations and equipment and lack broad external validity. Secondly, the study focused solely on common gastric focal lesions. Future work should include diffuse mucosal conditions, such as gastritis, atrophy, and intestinal metaplasia, as well as rare conditions such as post-gastrectomy patients, to better evaluate the system's broader applicability and provide more comprehensive clinical guidance. Thirdly, the AI-assisted system improved report completeness and consistency compared with the conventional semi-structured workflow (fixed regional headers with manual descriptive input by endoscopists). However, it was not compared with fully structured reporting systems, and its relative advantage in such settings remains to be determined. Cost-effectiveness and potential workflow disruption were not assessed in this study, but should be systematically investigated to determine the feasibility of real-world implementation. Finally, although the prospective observational validation provided preliminary empirical data, future large-scale prospective studies or randomized controlled trials are needed to more accurately evaluate the clinical applicability of the reporting system and minimize potential biases.

Conclusion

In this study, we validated an AI reporting system for gastric focal lesions. The results showed that the AI system enhances the completeness and accuracy of reports generated by endoscopists. This system is anticipated to reduce human omissions and provide more consistent reporting, thereby aiding in the standardization of medical information.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00464-025-12410-z.

Author contributions HGY, HY, and ZHD conceived and designed the study; HY and ZHD collected and analyzed the datasets; HY and XT drafted the manuscript; ZHD critically revised the article; HGY sponsored the study; all authors reviewed and approved the final manuscript for submission. We ensured that all the authors had access to all the raw datasets. All authors were involved in data acquisition, general design of the trial, interpretation of the data, and critical revision of the manuscript.

Funding This work was partly supported by the Key Research and Development Program of Hubei Province (grant no. 2023BCB153, to Honggang Yu), the College-enterprise Deepening Reform Project of Wuhan University (to Honggang Yu), National Natural Science Foundation of China-Youth Science Fund (grant no. 82202257, to Lianlian Wu), Natural Science Foundation of Wuhan (2025040601020197 to Zehua Dong), Hubei Provincial Key Laboratory Open Project (2024KFZ005 to Zehua Dong) and the National Key Research and Development Program of China (grant no. 2022YFC2505100, to Lianlian Wu, Wei Zhou).

Declarations

Conflict of interest Hang You, Xiao Tao, Zehua Dong, Yanxia Li, Li Huang, Jia Li,. Hongliu Du, Mei Deng, Zhifeng Wu, Xia Tan, Ting Yang, Jun Liu, and Honggang Yu have no conflict of interest or financial ties to disclose.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Bisschops R, Areia M, Coron E, Dobru D, Kaskas B, Kuvaev R et al (2016) Performance measures for upper gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. Endoscopy 9:843–864

2. Saftoiu A, Hassan C, Areia M, Bhutani MS, Bisschops R, Bories E et al (2020) Role of gastrointestinal endoscopy in the screening of digestive tract cancers in Europe: European Society of Gastrointestinal Endoscopy (ESGE) position statement. Endoscopy 4:293–304
3. Bretthauer M, Aabakken L, Dekker E, Kaminski MF, Rosch T, Hultcrantz R et al (2016) Requirements and standards facilitating quality improvement for reporting systems in gastrointestinal endoscopy: European Society of Gastrointestinal Endoscopy (ESGE) position statement. Endoscopy 3:291–294
4. Lieberman D, Nadel M, Smith RA, Atkin W, Duggirala SB, Fletcher R et al (2007) Standardized colonoscopy reporting and data system: report of the Quality Assurance Task Group of the National Colorectal Cancer Roundtable. Gastrointest Endosc 6:757–766
5. Boys JA, Azadgoli B, Martinez M, Oh DS, Hagen JA, DeMeester SR (2021) Adequacy of EGD reporting: a review of 100 reports from 100 endoscopists. J Gastrointest Surg 5:1117–1123
6. Zhang L, Lu Z, Yao L, Dong Z, Zhou W, He C et al (2023) Effect of a deep learning-based automatic upper GI endoscopic reporting system: a randomized crossover study (with video). Gastrointest Endosc. 98(2):181–90
7. Dong Z, Wu L, Mu G, Zhou W, Li Y, Shi Z et al (2022) A deep learning-based system for real-time image reporting during esophagogastroduodenoscopy: a multicenter study. Endoscopy 8:771–777
8. Faigel DO, Pike IM, Baron TH, Chak A, Cohen J, Deal SE et al (2006) Quality indicators for gastrointestinal endoscopic procedures: an introduction. Gastrointest Endosc 4 Suppl:S3-9
9. Cohen J, Safdi MA, Deal SE, Baron TH, Chak A, Hoffman B et al (2006) Quality indicators for esophagogastroduodenoscopy. Am J Gastroenterol 101(4):886–891
10. Armstrong D, Barkun A, Bridges R, Carter R, de Gara C, Dube C et al (2012) Canadian Association of Gastroenterology consensus guidelines on safety and quality indicators in endoscopy. Can J Gastroenterol 1:17–31
11. Raftopoulos SC, Segarajasingam DS, Burke V, Ee HC, Yusoff IF (2010) A cohort study of missed and new cancers after esophagogastroduodenoscopy. Am J Gastroenterol 6:1292–1297
12. Telford JJ, Enns RA (2010) Endoscopic missed rates of upper gastrointestinal cancers: parallels with colonoscopy. Am J Gastroenterol 105(6):1298–1300
13. Ezoe Y, Muto M, Horimatsu T, Minashi K, Yano T, Sano Y et al (2010) Magnifying narrow-band imaging versus magnifying white-light imaging for the differential diagnosis of gastric small depressive lesions: a prospective study. Gastrointest Endosc 3:477–484
14. Kato M, Kaise M, Yonezawa J, Toyoizumi H, Yoshimura N, Yoshida Y, Kawamura M, Tajiri H (2010) Magnifying endoscopy with narrow-band imaging achieves superior accuracy in the differential diagnosis of superficial gastric lesions identified with white-light endoscopy: a prospective study. Gastrointest Endosc 72(3):523–529. https://doi.org/10.1016/j.gie.2010.04.041
15. Zhang Q, Wang F, Chen ZY, Wang Z, Zhi FC, Liu SD et al (2016) Comparison of the diagnostic efficacy of white light endoscopy and magnifying endoscopy with narrow band imaging for early gastric cancer: a meta-analysis. Gastric Cancer 19(2):543–552
16. Dong Z, Tao X, Du H, Wang J, Huang L, He C et al (2023) Exploring the challenge of early gastric cancer diagnostic AI system face in multiple centers and its potential solutions. J Gastroenterol 58(10):978–989
17. Dong Z, Wang J, Li Y, Deng Y, Zhou W, Zeng X et al (2023) Explainable artificial intelligence incorporated with domain knowledge diagnosing early gastric neoplasms under white light endoscopy. NPJ Digit Med 6(1):64
18. Lieberman DA, Faigel DO, Logan JR, Mattek N, Holub J, Eisen G, Morris C, Smith R, Nadel M (2009) Assessment of the quality of colonoscopy reports: results from a multicenter consortium. Gastrointest Endosc 69(3 Pt 2):645–653. https://doi.org/10.1016/j.gie.2008.08.034
19. He X, Wu L, Dong Z, Gong D, Jiang X, Zhang H et al (2002) Real-time use of artificial intelligence for diagnosing early gastric cancer by magnifying image-enhanced endoscopy: a multicenter diagnostic study (with videos). Gastrointest Endosc. 95(4):671–8
20. Luo H, Xu G, Li C, He L, Luo L, Wang Z, Jing B, Deng Y, Jin Y, Li Y, Li B, Tan W, He C, Seeruttun SR, Wu Q, Huang J, Huang D-W, Chen B, Lin S-B, Chen Q-M, Yuan C-M, Chen H-X, Pu H-Y, Zhou F, He Y, Xu R-H (2019) Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. Lancet Oncol 20(12):1645–1654. https://doi.org/10.1016/S1470-2045(19)30637-0
21. Niikura R, Aoki T, Shichijo S, Yamada A, Kawahara T, Kato Y et al (2022) Artificial intelligence versus expert endoscopists for diagnosis of gastric cancer in patients who have undergone upper gastrointestinal endoscopy. Endoscopy 54(8):780–784
22. Tang D, Ni M, Zheng C, Ding X, Zhang N, Yang T et al (2022) A deep learning-based model improves diagnosis of early gastric cancer under narrow band imaging endoscopy. Surg Endosc 36(10):7800–7810
23. Koh FH, Ladlad J, Teo EK, Lin CL, Foo FJ (2023) Real-time artificial intelligence (AI)-aided endoscopy improves adenoma detection rates even in experienced endoscopists: a cohort study in Singapore. Surg Endosc 37(1):165–171
24. Lu Z, Zhang L, Yao L, Gong D, Wu L, Xia M, Zhang J, Zhou W, Huang Xu, He C, Wu H, Zhang C, Li X, Yu H (2023) Assessment of the role of artificial intelligence in the association between time of day and colonoscopy quality. JAMA Netw Open 6(1):e2253840. https://doi.org/10.1001/jamanetworkopen.2022.53840
25. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P (2024) Heterogeneity and predictors of the effects of AI assistance on radiologists. Nat Med 30(3):837–849

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.