Development and Validation of an Automatic Image-Recognition Endoscopic Report Generation System: A Multicenter Study

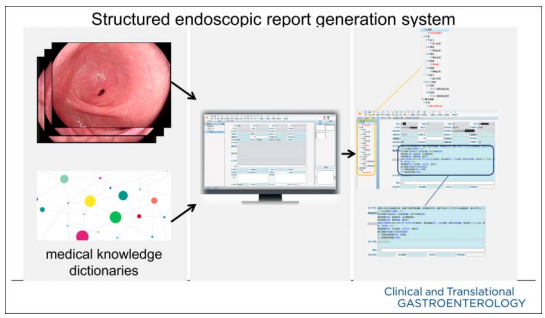
Jun-yan Qu, MD[1,2,3], Zhen Li, MD, PhD[1,2,3], Jing-ran Su, MD[1,2,3], Ming-jun Ma, MD[1,2,3], Chang-qin Xu, MD, PhD[4], Ai-jun Zhang, MD[5], Cheng-xia Liu, MD, PhD[6], Hai-peng Yuan, MD, PhD[7], Yan-liu Chu, MD, PhD[8], Cui-cui Lang, MD[9], Liu-ye Huang, MD[10], Lin Lu, MD, PhD[11], Yan-qing Li, MD, PhD[1,2,3] and Xiu-li Zuo, MD, PhD[1,2,3]

INTRODUCTION: **Conventional gastrointestinal (GI) endoscopy reports written by physicians are time consuming and might have obvious heterogeneity or omissions, impairing the efficiency and multicenter consultation potential. We aimed to develop and validate an image recognition–based structured report generation system (ISRGS) through a multicenter database and to assess its diagnostic performance.**

METHODS: **First, we developed and evaluated an ISRGS combining real-time video capture, site identification, lesion detection, subcharacteristics analysis, and structured report generation. White light and chromoendoscopy images from patients with GI lesions were eligible for study inclusion. A total of 46,987 images from 9 tertiary hospitals were used to train, validate, and multicenter test (6:2:2). Moreover, 5,699 images were prospectively enrolled from Qilu Hospital of Shandong University to further assess the system in a prospective test set. The primary outcome was the diagnosis performance of GI lesions in multicenter and prospective tests.**

RESULTS: **The overall accuracy in identifying early esophageal cancer, early gastric cancer, early colorectal cancer, esophageal varices, reflux esophagitis, Barrett's esophagus, chronic atrophic gastritis, gastric ulcer, colorectal polyp, and ulcerative colitis was 0.8841 (95% confidence interval, 0.8775–0.8904) and 0.8965 (0.8883–0.9041) in multicenter and prospective tests, respectively. The accuracy of cecum and upper GI site identification were 0.9978 (0.9969–0.9984) and 0.8513 (0.8399–0.8620), respectively. The accuracy of staining discrimination was 0.9489 (0.9396–0.9568). The relative error of size measurement was 4.04% (range 0.75%–7.39%).**

DISCUSSION: **ISRGS is a reliable computer-aided endoscopic report generation system that might assist endoscopists working at various hospital levels to generate standardized and accurate endoscopy reports (http://links.lww.com/CTG/A485).**



[1]Department of Gastroenterology, Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan, China; [2]Laboratory of Translational Gastroenterology, Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan, China; [3]Robot Engineering Laboratory for Precise Diagnosis and Therapy of GI Tumor, Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan, China; [4]Department of Gastroenterology, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, China; [5]Department of Gastroenterology, Qilu Hospital of Shandong University (Qingdao), Qingdao, China; [6]Department of Gastroenterology, Binzhou Medical University Hospital, Binzhou, China; [7]Department of Gastroenterology, Taian City Central Hospital, Taian, China; [8]Department of Gastroenterology, Weihai Municipal Hospital, Weihai, China; [9]Department of Gastroenterology, Liaocheng People's Hospital, Liaocheng, China; [10]Department of Gastroenterology, Yantai Yuhuangding Hospital, Yantai, China; [11]Department of Gastroenterology, Linyi People's Hospital, Linyi, China. **Correspondence:** Xiu-li Zuo, MD, PhD. E-mail: zuoxiuli@sdu.edu.cn. Yan-qing Li, MD, PhD. E-mail: liyanqing@sdu.edu.cn.
Received August 28, 2020; accepted November 5, 2020; published online December 22, 2020

SUPPLEMENTARY MATERIAL accompanies this paper at http://links.lww.com/CTG/A480, http://links.lww.com/CTG/A481, http://links.lww.com/CTG/A482, http://links.lww.com/CTG/A483, http://links.lww.com/CTG/A484, http://links.lww.com/CTG/A485

Clinical and Translational Gastroenterology 2021;12:e00282. https://doi.org/10.14309/ctg.0000000000000282

INTRODUCTION

The wide application of gastrointestinal (GI) endoscopy makes it possible to diagnose and treat digestive diseases in the early stage, improving the prognosis of patients (1–3). In the conventional GI endoscopy report system, the diagnosis is made by endoscopists after subjective assessment of the endoscopic procedure (4–6), and a report is then generated manually by a physician through the computer mouse, keyboard, and other input devices.

However, errors seem to be unavoidable during this manual operation (7), introducing large interoperator and intraoperator variability because of the heterogeneity in experience, working habits, and state of endoscopists, impeding the accuracy, and standardization of the report (8–12). In addition, time involved reduces the efficiency of endoscopic examinations (13,14).

In recent years, major advances in artificial intelligence (AI) have occurred, especially in the recognition and characterization of representative visual data in the GI tract. It has been reported that AI enables detection or diagnosis of several GI neoplasms (15–17). Furthermore, a deep learning convolutional neural network (CNN)-based model has been shown to assist in the identification of small bowel abnormalities (18). Despite these findings, the performance of AI in multitarget recognition of 10 major GI diseases simultaneously and the generation of a standardized text report remains unclear.

Therefore, we developed an image recognition–based structured report generation system (ISRGS), which works through deep learning CNN models combining real-time video capture, site identification, diagnosis of GI lesions and their subcharacteristics analysis, and structured report generation. The primary objective of the study was to test the diagnostic performance of ISRGS using multicenter and prospective data sets.

METHODS

Study design

We developed and evaluated an ISRGS for real-time video capture, site identification, diagnosis of GI lesions, and subcharacteristics analysis of lesions through training, validation, multicenter, and prospective tests.

Image preparation for data sets and quality control

Two parts of images were used for the development and evaluation of ISRGS: (i) from July to October 2019, GI endoscopic images were retrospectively collected from 9 tertiary hospitals across China: Qilu Hospital of Shandong University (QHSU), Shandong Provincial Hospital Affiliated to Shandong First Medical University, QHSU (Qingdao), Liaocheng People's Hospital, Linyi People's Hospital, Weihai Municipal Hospital, Taian City Central Hospital, Binzhou Medical University Hospital, and Yantai Yuhuangding Hospital for the training, validation, and multicenter test of the models in a 6:2:2 ratio; (ii) from November 2019 to December 2019, video capture and diagnosis modules were applied to the original endoscopy monitors (EPK-i7000, Pentax, Tokyo, Japan) in QHSU for real-time video-based analysis, which was set to process images at 10 frames

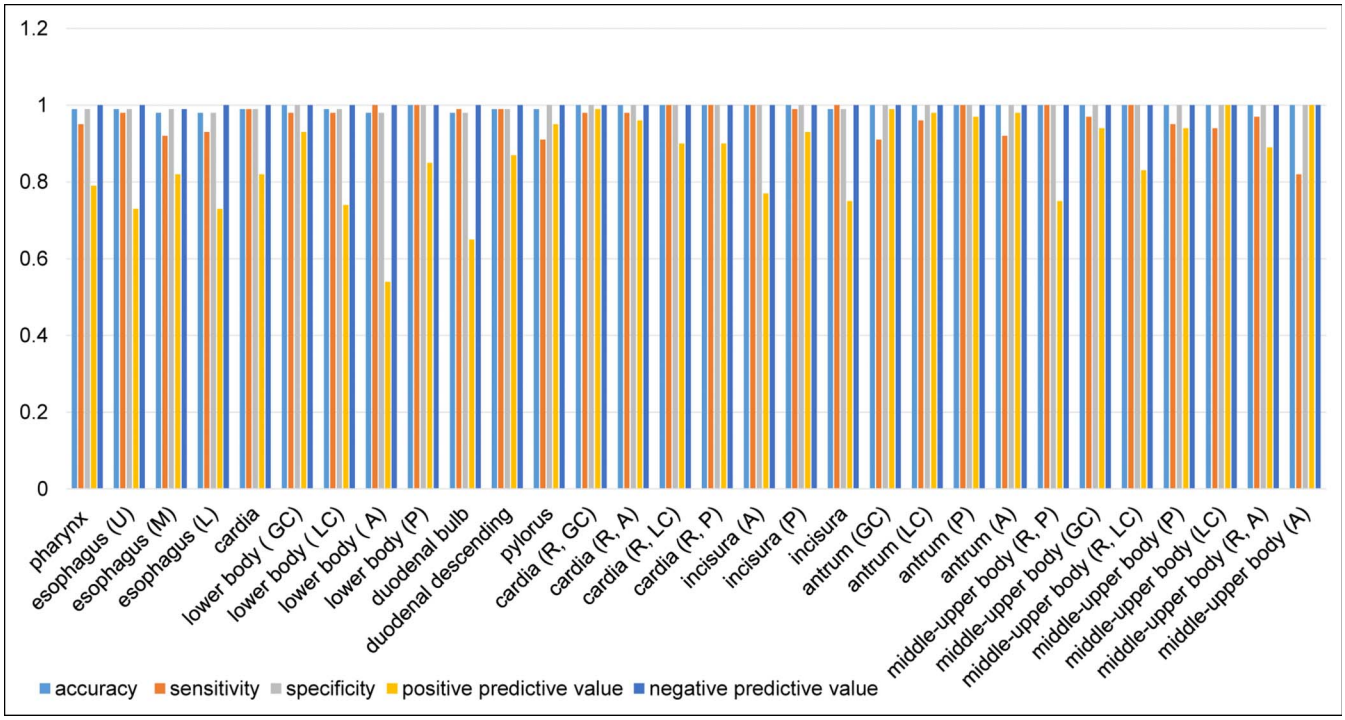


Figure 1. Test result for the upper GI site identification model. A, anterior wall; GC, greater curvature; GI, gastrointestinal; L, lower; LC, lesser curvature; M, middle; P, posterior wall; R, retroflex view; U, upper.

per second. Images from consecutive participants (with at least 1 of the 10 GI diseases) were enrolled into a prospective data set.

Endoscopic images were captured by endoscopes from different vendors in multiple centers, including GIF-H260Z, GIF-HQ290, GIF-XQ260, GIF-Q260, GIF-H260, CF-H290, CF-HQ290, CF-H260, CF-HQ260, CF-Q260, PCF-Q260, PCF-PQ260, Olympus, Japan; EG-2990i, EG29-i10, EG27-i10, EC38-i10, EC-3490, EC-3870, EC-3890, Pentax, Japan; and EG-580RD, EC-L590, Fujifilm, Japan.

Normal and abnormal images (including: early esophageal cancer, early gastric cancer, early colorectal cancer, esophageal varices, reflux esophagitis, Barrett's esophagus, chronic atrophic gastritis, gastric ulcer, colorectal polyp, and ulcerative colitis) with white light and chromoendoscopy were selected. The labeling and delineation data finalized by experienced endoscopists were regarded as the gold standard in this study.

Images were assigned to 4 experienced endoscopists with a minimum of 5 years of experience. They assessed the images quality, and lesions were labeled and delineated. Then, each image was assessed by another 2 highly experienced endoscopists for quality control. Differences between the reviewers were resolved through discussion.

Development of the ISRGS algorithm

The ISRGS consists of the following 7 submodules: image acquisition, site identification, and lesion detection; 3 subcharacteristics (staining recognition, effective biopsy judgment, and lesion size estimation) analysis; and structured report generation. The input of ISRGS was GI endoscopic images. The outputs in the generated structured report consisted of the sites and diagnosis of lesions and subcharacteristics of the procedure.

The InceptionResnet v2 and MobileNet v3 neural networks were applied to the site identification module for the upper GI tract and the cecum, respectively. Endoscopic upper GI tract data were divided into 30 categories based on the systematic screening protocol of GI tract (19).

As a multitask learning architecture, the You Only Look Once (YOLO) v3 neural network was used in the lesion detection module because of its high diagnostic accuracy and fast detection speed, meeting the needs of real-time monitoring of endoscopy. The size of input images was 416 × 416 × 3 pixels. The output provides 3 bounding boxes of different sizes, which can detect targets with different sizes in the image. Each output predicted whether there were targets in the bounding box, the box, and the classification of prediction. The loss function of YOLO v3 is generally consistent with that of YOLO v1.

SSD300 and C3D algorithms were used for the judgment of effective biopsy processes, defined as the continuous appearance of biopsy forceps in the endoscopic field of vision (from the opening of biopsy forceps to the pulling of biopsy forceps) and the appearance of biopsy scars (often with bleeding) on the inner

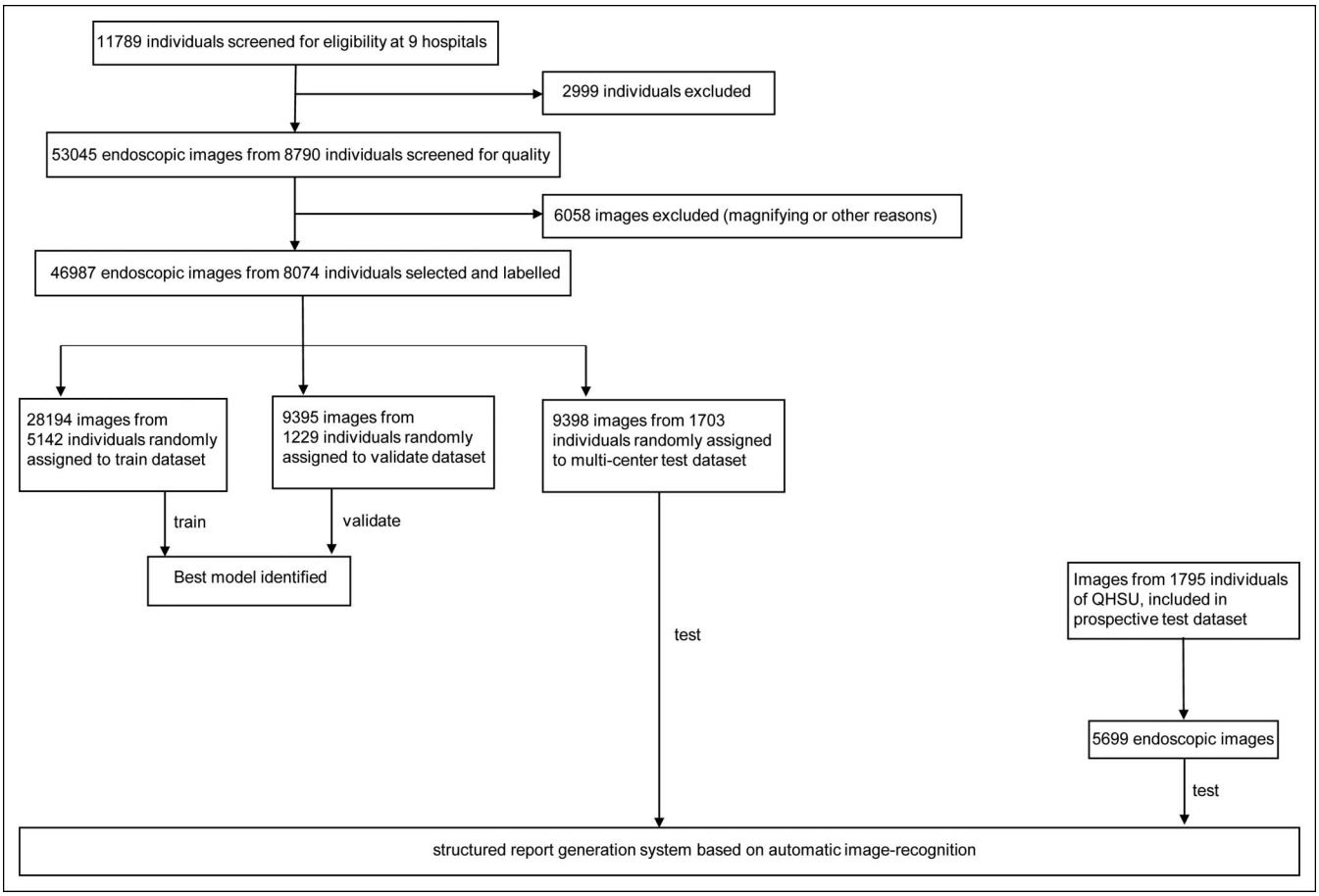


Figure 2. Flowchart for the training, validation, and test of the ISRGS. ISRGS, image-recognition-based structured report generation system; QHSU, Qilu Hospital of Shandong University.

Table 1. Baseline characteristics

	Training set	Validation set	Multicenter test set	Prospective test set
Age (yr), mean (SD)	56.89 (14.75)	51.18 (12.85)	55.61 (13.36)	52.56 (12.64)
Sex				
Male, n (%)	2,719 (52.88)	592 (48.17)	832 (48.85)	954 (53.15)
Female, n (%)	2,423 (47.12)	637 (51.83)	871 (51.15)	805 (44.85)
Early esophageal cancer, n (%)	1,512 (5.36)	504 (5.36)	504 (5.36)	208 (3.65)
Reflux esophagitis, n (%)	1,692 (6.00)	565 (6.01)	564 (6.00)	262 (4.60)
Esophageal varices, n (%)	1,539 (5.46)	513 (5.46)	513 (5.46)	306 (5.37)
Barrett's esophagus, n (%)	1,698 (6.02)	566 (6.02)	566 (6.02)	320 (5.62)
Early gastric cancer, n (%)	1,524 (5.41)	508 (5.41)	508 (5.41)	108 (1.90)
Gastric ulcer, n (%)	2,649 (9.40)	884 (9.41)	883 (9.40)	195 (3.42)
Chronic atrophic gastritis, n (%)	1,551 (5.50)	517 (5.50)	517 (5.50)	405 (7.11)
Early colorectal cancer, n (%)	2,589 (9.18)	862 (9.18)	863 (9.18)	228 (4.00)
Colorectal polyp, n (%)	4,119 (14.61)	1,372 (14.60)	1,373 (14.61)	559 (9.81)
Ulcerative colitis, n (%)	1,821 (6.46)	606 (6.45)	607 (6.46)	279 (4.90)
No disease	7,500 (26.60)	2,498 (26.59)	2,500 (26.60)	2,829 (49.64)

wall of the digestive tract (see Figure, Supplementary Digital Content 1, http://links.lww.com/CTG/A480). SSD300 (the target detection model) continuously tracked and located the position of the biopsy forceps and then captured a video of the biopsy scene through logical judgment. Furthermore, C3D (the scene recognition model) determined whether the video scene was an effective biopsy process.

For staining recognition, we used MobileNet v3 to determine whether the patients were stained with dye or optical staining techniques, including iodine, methylene blue, indigo carmine, narrow band imaging, optical enhancement and flexible spectral imaging color enhancement. To estimate lesion size, the endoscopy water column was used as the standard reference for lesion measurement, and the Mask-Region-based-Convolutional Neural Networks was used to identify the width of the junction between the forward water column and the GI mucosa (as the standard measurement scale). Lesion size was calculated according to the image reference width change (see Figure, Supplementary Digital Content 2, http://links.lww.com/CTG/A481).

According to the results of site and lesion recognition, combined with the medical knowledge dictionaries, the corresponding description text was generated and added to the structured template to obtain a structured report (see Figure, Supplementary Digital Content 3, http://links.lww.com/CTG/A482). In this module, the Image Caption model was used for semantic understanding of the current video frame to obtain a substance naming description, whereas encoder–decoder model was used to generate natural language description text.

Training and validation of the ISRGS

In the training phase, we retrained the submodules using the labeled data sets and updated the parameters of all layers by continuous optimization. The training was stopped when the total loss of the models was stable on independent validation data sets.

In the site identification module, 19,645 endoscopic images consisting of the cecum and other colorectal sites were included in the training, validation, and testing data sets for the cecum identification model, whereas 4,000 endoscopic images captured from real-time videos of gastroscopy were used in the upper GI site identification model. Testing results showed that our AI model could discriminate cecum from other colorectal sites with an accuracy of 0.9978 (95% confidence interval [CI], 0.9969–0.9984). For the upper GI site identification model, the total accuracy was 0.8513 (95% CI, 0.8399–0.8620) (Figure 1).

For the lesion detection module, the training platform of YOLO v3 was Darknet. The model was trained for 453 epochs with a batch size of 64. We used a trick training network (the method of dynamic learning rate), with the initial learning rate set as 0.001. The network learned with AdamOptimizer to optimize the total loss of the model during the training process. To ensure that at least 500 representative images for each category, 46,987 labeled and delineated endoscopic images were selected in the training, validation, and multicenter test set. We evaluated the diagnostic accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of ISRGS in identifying 10 GI lesions. Total accuracy was measured as the number of correctly diagnostic images (by algorithm) divided by the total number. Sensitivity and specificity were defined as the proportion of the gold standard confirmed lesions and the proportion of negative controls, respectively, which were correctly identified by the algorithm. Positive and negative predictive values were defined as the proportion of algorithm-suggested lesions that were true lesions and the proportion of algorithm-suggested nontarget lesion images that were true nontarget lesions, respectively.

In the effective biopsy judgment module, we trained the SSD300 and C3D algorithms. We labeled 400 biopsy forceps images and used the ImageNet pretraining model to train SSD30. The size of the input image was $300 \times 300 \times 3$ pixels. For C3D, we first obtained 100 effective biopsy and nonbiopsy videos during endoscopic procedures using video editing tools. Based on the video clips, we obtained the video frames using FFmpeg in order

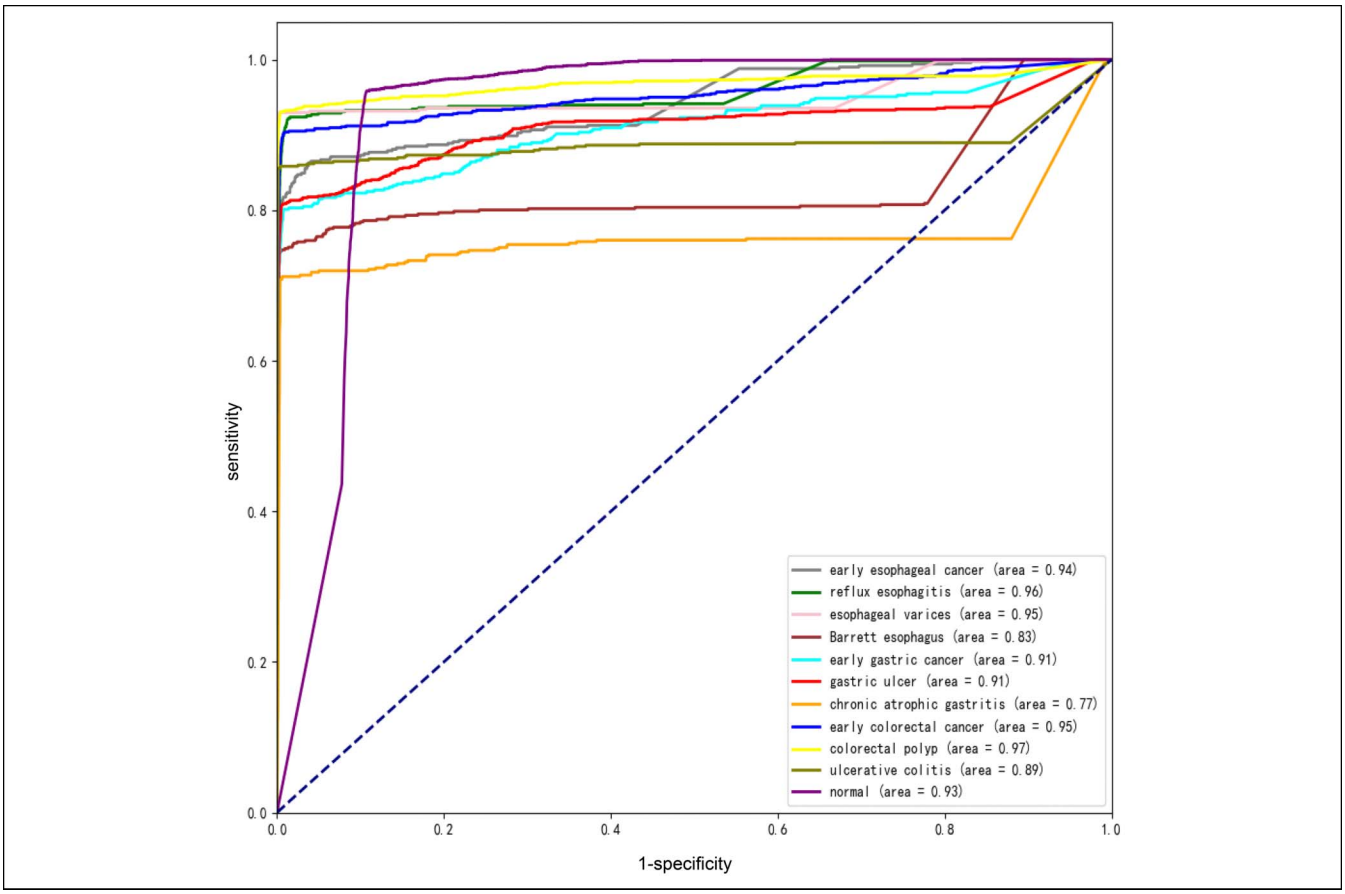


Figure 3. Receiver operating characteristic curves of multicenter test set in diagnosing GI diseases. GI, gastrointestinal.

and then sampled 16 images at equal intervals. In this way, we obtained 16 images for each video clip, removed the black edge removal and scaling processing for these 16 images, and finally obtained the 112 × 112 × 3 × 16 video sequence image matrix. Another 40 videos were used to test the module. For the total 48 effective biopsy processes of the 20 biopsy videos, an incorrect identification was reported for the occlusion of the view, while there was no misidentification among the other non-biopsy 20 videos.

In addition, another 2,523 chromoendoscopy images were selected and labeled to test the staining recognition module, with a total accuracy of 0.9489 (95% CI, 0.9396–0.9568). The model was trained for 300 epochs with a batch size of 128. We also used a trick training network, with the initial learning rate set as 0.001. The network learned with AdamOptimizer to optimize the total loss of the model during the training process.

To estimate lesion size, 1,356 images were selected to train and validate the algorithm. In April 2019, we used circular scales with different diameters to imitate GI lesions in the porcine stomach to test the diagnostic performance of this module. We found that the accuracy of measured lesions within 5 cm of vertical distance and within 4 cm of lesion size was the most reliable, owing to the barrel distortion of endoscopic images and the fluid characteristics of the water jet. When the water jet was sprayed vertically on discs of different diameters *in vitro* to simulate the porcine stomach focus measurement, the mean relative error was 4.04% (range 0.75%–7.39%).

Testing the ISRGS

First, we tested the performance of the ISRGS in the diagnosis of 10 GI diseases in a multicenter test data set. We then prospectively enrolled images from consecutive participants with 1 of the 10 GI diseases for further performance assessment in clinical practice. There was no patient overlap among the data sets.

Statistical analysis

We used the receiver operating characteristic (ROC) curve to show the diagnostic ability of the deep learning algorithm, in which the probability of algorithm prediction was introduced to obtain the tradeoff between sensitivity and specificity for each disease category. A larger area under the ROC curve indicated better diagnostic performance. Statistical analyses were performed using SPSS 25.0 software for Windows (IBM, Armonk, NY). This study was approved by the relevant independent institutional review boards of each participating hospital, and any personally identifying information was omitted. Informed consent was exempted by the institutional review boards of the participating hospitals.

RESULTS

Baseline characteristics of data set

Between July 2019 and October 2019, we retrospectively accessed 11,789 individuals from 9 tertiary hospitals for eligibility, of which 2,999 individuals with normal GI mucosa were excluded. Then, 53,045 endoscopic images from 8,790 individuals were

Table 2. The performance of ISRGS on diagnosis of GI diseases in multicenter test set

	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Early esophageal cancer	0.9910 (0.9889–0.9927)	0.8591 (0.8250–0.8877)	0.9984 (0.9973–0.9991)	0.9687 (0.9467–0.9821)	0.9921 (0.9899–0.9938)
Reflux esophagitis	0.9896 (0.9873–0.9915)	0.9220 (0.8959–0.9421)	0.9939 (0.9920–0.9954)	0.9059 (0.8783–0.9279)	0.9950 (0.9932–0.9963)
Esophageal varices	0.9960 (0.9945–0.9971)	0.9298 (0.9033–0.9497)	0.9998 (0.9991–1.0000)	0.9958 (0.9833–0.9993)	0.9960 (0.9944–0.9971)
Barrett's esophagus	0.9823 (0.9794–0.9848)	0.7438 (0.7054–0.7789)	0.9976 (0.9963–0.9985)	0.9525 (0.9271–0.9696)	0.9838 (0.9809–0.9863)
Early gastric cancer	0.9818 (0.9789–0.9843)	0.8012 (0.7632–0.8345)	0.9921 (0.9900–0.9938)	0.8532 (0.8175–0.8831)	0.9887 (0.9862–0.9907)
Gastric ulcer	0.9765 (0.9732–0.9794)	0.8075 (0.7796–0.8327)	0.9940 (0.9921–0.9955)	0.9332 (0.9126–0.9494)	0.9803 (0.9771–0.9831)
Chronic atrophic gastritis	0.9804 (0.9774–0.9830)	0.7079 (0.6663–0.7464)	0.9963 (0.9947–0.9974)	0.9173 (0.8847–0.9416)	0.9832 (0.9803–0.9857)
Early colorectal cancer	0.9841 (0.9814–0.9864)	0.9027 (0.8804–0.9212)	0.9924 (0.9902–0.9941)	0.9230 (0.9024–0.9396)	0.9902 (0.9878–0.9921)
Colorectal polyp	0.9862 (0.9836–0.9884)	0.9308 (0.9158–0.9434)	0.9956 (0.9939–0.9969)	0.9733 (0.9627–0.9811)	0.9882 (0.9856–0.9904)
Ulcerative colitis	0.9901 (0.9879–0.9919)	0.8550 (0.8239–0.8816)	0.9994 (0.9986–0.9998)	0.9905 (0.9765–0.9965)	0.9901 (0.9877–0.9920)

CI, confidence interval; GI, gastrointestinal; ISRGS, image recognition–based structured report generation system; NPV, negative predictive value; PPV, positive predictive value.

screened for quality, and 6,058 images with magnifying or other reasons were excluded. As a result, 46,987 endoscopic images from 8,074 individuals were labeled and selected for training, validation, and internal test data sets in the ratio of 6:2:2. Between November 2019 and December 2019, 5,699 video-based images from 1,795 consecutive participants were enrolled into a prospective data set (Figure 2). Detailed baseline characteristics and constituent ratios of the GI diseases for the data sets are summarized in Table 1.

The performance of ISRGS in diagnosis of GI diseases of multi-center test set. High area under the ROC curve values were observed (Figure 3). The optimal threshold value of the ISRGS was derived from the ROC curve according to the Youden index method.

The overall accuracy of the ISRGS for identifying the target 10 diseases was 0.8841 (95% CI, 0.8775–0.8904). Table 2 displays the excellent performance of ISRGS in detecting 10 major GI diseases in the multicenter test set, with accuracies ranging from 0.9765 (95% CI, 0.9732–0.9794) to 0.9960 (95% CI, 0.9945–0.9971). See Table, Supplementary Digital Content 4, http://links.lww.com/CTG/A483 for the confusion matrix of the prospective test phase.

The performance of ISRGS in diagnosis of GI diseases of prospective test set

Figure 4 shows the ROC curve. The overall accuracy for the prospective test set was 0.8965 (95% CI, 0.8883–0.9041). The accuracy ranged from 0.9810 (95% CI, 0.9771–0.9842) to 0.9919 (95% CI, 0.9892–0.9939) (Table 3). See Table, Supplementary Digital Content 5, http://links.lww.com/CTG/A484 for the confusion matrix of the prospective test phase.

DISCUSSION

ISGRS was integrated into 7 submodules and verified with more than 52,686 images from 9,869 participants from 9 tertiary hospitals. It should be noted that, as a universal, standardized, and efficient endoscopy-assisted system, the ISGRS was capable of correctly diagnosing and framing out most common GI diseases, including early esophageal, gastric, and colorectal cancer, esophageal varices, reflux esophagitis, Barrett's esophagus, chronic atrophic gastritis, gastric ulcer, colorectal polyp, and ulcerative colitis.

This study has multiple strengths. First, the ISRGS is the first AI system able to diagnose multiple types of GI diseases in real-time. Second, to the best of our knowledge, the automatic structured report generation module is also the first image text conversion system, which helps to standardize diagnosis description, reduce labor resources, and improve endoscopy efficiency. Another highlight of this system is that the generalization of the ISRGS was verified by multicenter large sample data generated by different endoscopy vendors.

Luo et al. (17) reported a real-time AI module for detection of upper GI cancer. Ding et al. (18) developed a CNN-based algorithm to identify abnormalities in small bowel capsule endoscopy images. However, both showed binary results, which were not sufficient to identify multiple abnormalities in clinical practice. In

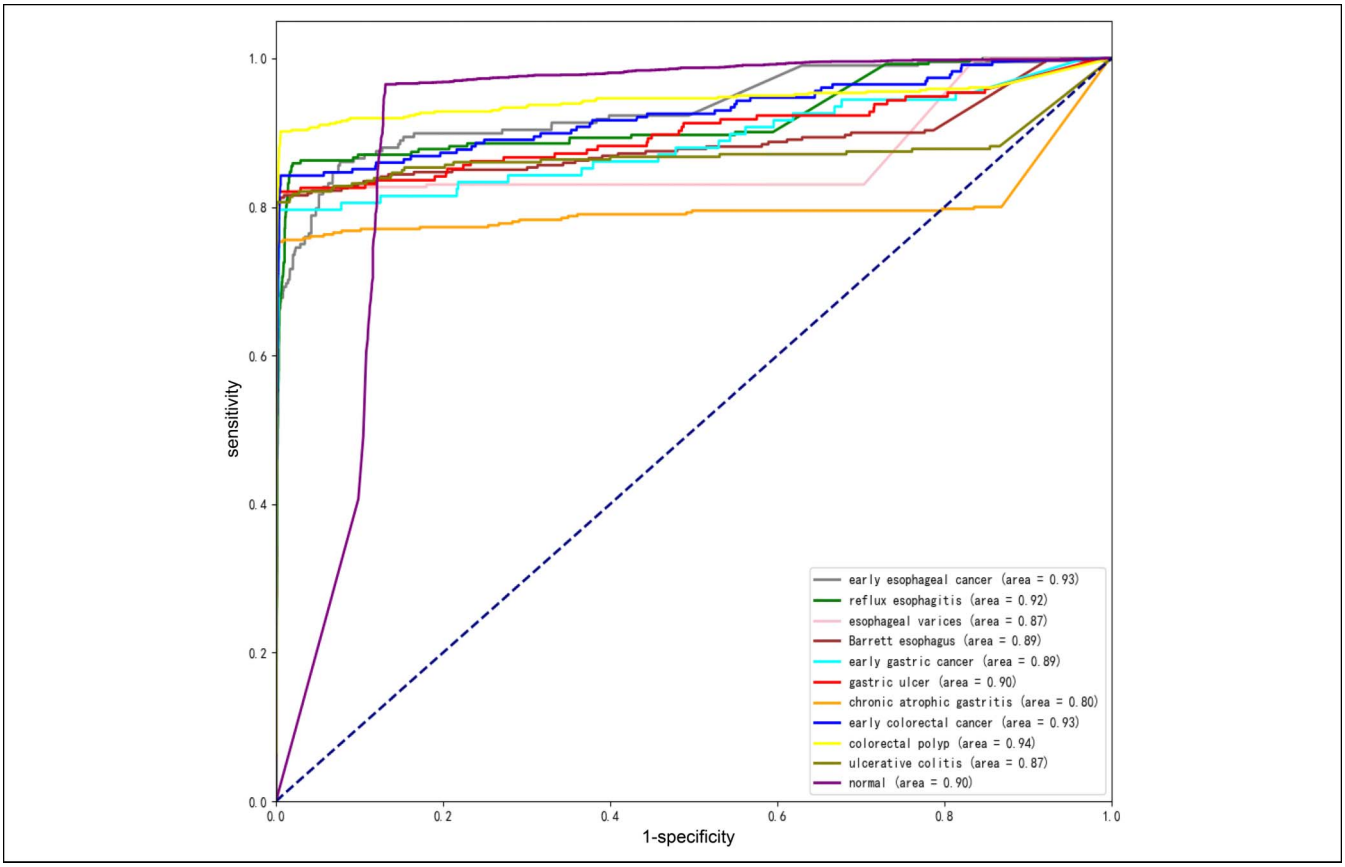


Figure 4. Receiver operating characteristic curves of prospective test set in diagnosing GI diseases. GI, gastrointestinal.

this study, the multitask learning architecture YOLO v3 neural network was used to identify the lesions, in which intersection over union (IoU) was introduced for performance assessment. IoU is defined as the ratio of the intersection and union of the AI-predicted border and gold standard border. In this study, IoU was set as 10%. In detail, in a picture containing the target lesion, when the algorithm correctly predicted the lesion label and at least 1 IoU was more than 10%, the prediction was regarded as true positive. Advanced GI cancer was not assessed in the system because of its obvious characteristics, which can be correctly recognized even by trainees. If advanced GI cancer was included in the data set, the accuracy of the ISRGS in the detection of GI disease would improve.

Chromoendoscopy facilitates the differentiation between cancerous and noncancerous lesions (20–23). Although guidelines recommend image-enhanced endoscopy (IEE) including dye-based IEE and equipment-based IEE to improve the detection rate of GI neoplasia, doctors might forget or omit it in real-world practice. In addition, omissions might occur in the IEE method during report generation, which can lead to massive data being missing (24,25). Meanwhile, endoscopic assessment of lesion size is also subjective in our current situation. Using the subcharacteristic assessment modules of the ISGRS, both IEE type and the lesion size could be accurately recorded in the endoscopic report in real-time. Moreover, effective biopsy movement could also be recognized by the ISGRS and automatically included in the report generation system, helping to construct a complete endoscopy result sheet.

In this study, when AI identifies a lesion in real-time, it will analyze about the stability of the view, which takes around 30 ms. Furthermore, endoscopic photographs with no artifact will be captured. For colonoscopy, AI could also improve polyp and adenoma detection rates and calculate withdrawal time by accurately recognizing the cecum and the *in vivo* and *ex vivo* timepoint (26). In this study, the recognizing speed of site using the ISGRS is 10 frames per second, and the process of disease recognition takes approximately 70 ms. The final recognition result is displayed on the ISGRS monitor, which is adjacent to the original endoscopy screen. There was only a latency of 500 to 600 ms during real-time video analysis.

There are also several limitations to this study. First, there is no true gold standard for some GI diseases in this study (such as esophageal varices, reflux esophagitis, and colorectal polyp); thus, potential bias might exist in deep learning. Second, the sensitivity for the diagnosis of early GI cancer ranged from 80% to 90%. Multiple categories of GI lesions and the misdiagnosis of other GI lesions might in part lead to the low sensitivity. Subgroup analysis showed that the overall sensitivity of early GI cancer was 87.2% and 84.6% in the multicenter and prospective test sets, respectively. In some degree, the output of the other diseases might be also useful, which could remind the endoscopist to observe carefully and avoid misdiagnosis. In addition, the histopathology was regarded as the gold standard in this study. Studies from China showed that the missing rate of early GI is nearly 17% to 59% (esophageal), 18% to 42% (gastric) and 9% to 27% (colorectal)

Table 3. The performance of ISRGS on diagnosis of GI diseases in prospective test set

	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Early esophageal cancer	0.9882 (0.9850–0.9907)	0.8462 (0.7882–0.8909)	0.9936 (0.9910–0.9955)	0.8341 (0.7754–0.8803)	0.9942 (0.9917–0.9959)
Reflux esophagitis	0.9860 (0.9826–0.9887)	0.8550 (0.8051–0.8941)	0.9923 (0.9895–0.9944)	0.8421 (0.7914–0.8827)	0.9930 (0.9903–0.9950)
Esophageal varices	0.9891 (0.9861–0.9915)	0.8105 (0.7610–0.8519)	0.9993 (0.9980–0.9998)	0.9841 (0.9571–0.9949)	0.9894 (0.9862–0.9918)
Barrett's esophagus	0.9856 (0.9822–0.9884)	0.8125 (0.7644–0.8529)	0.9959 (0.9937–0.9974)	0.9220 (0.8827–0.9494)	0.9889 (0.9857–0.9915)
Early gastric cancer	0.9919 (0.9892–0.9939)	0.7963 (0.7057–0.8653)	0.9957 (0.9935–0.9972)	0.7818 (0.6909–0.8526)	0.9961 (0.9939–0.9975)
Gastric ulcer	0.9911 (0.9883–0.9932)	0.8205 (0.7578–0.8702)	0.9971 (0.9952–0.9983)	0.9091 (0.8541–0.9455)	0.9937 (0.9911–0.9955)
Chronic atrophic gastritis	0.9810 (0.9771–0.9842)	0.7531 (0.7076–0.7937)	0.9985 (0.9969–0.9993)	0.9744 (0.9483–0.9881)	0.9814 (0.9774–0.9848)
Early colorectal cancer	0.9884 (0.9853–0.9909)	0.8421 (0.7867–0.8856)	0.9945 (0.9921–0.9962)	0.8649 (0.8111–0.9056)	0.9934 (0.9908–0.9953)
Colorectal polyp	0.9853 (0.9818–0.9881)	0.9016 (0.8731–0.9244)	0.9944 (0.9918–0.9961)	0.9456 (0.9219–0.9626)	0.9894 (0.9861–0.9919)
Ulcerative colitis	0.9896 (0.9866–0.9919)	0.8029 (0.7503–0.8469)	0.9993 (0.9980–0.9998)	0.9825 (0.9527–0.9944)	0.9899 (0.9868–0.9923)

CI, confidence interval; GI, gastrointestinal; ISRGS, image recognition–based structured report generation system; NPV, negative predictive value; PPV, positive predictive value.

in clinical practice (27–29). Further randomized controlled studies are warranted to fully evaluate the true benefit of this computer-aided system in automatic report generation. Third, despite the 3 subcharacteristics that can be evaluated by the ISGRS, there are still features in the structured report template that need to be selected by the endoscopist, such as the border, color changes, surface pattern. In the future, more techniques, such as natural language processing, might further improve the accuracy and operability of automatic endoscopic report generation system.

In summary, the ISGRS is the first AI-based endoscopic report generation system. It might serve as an effective tool to detect multiple GI lesions, assess certain subcharacteristics, and facilitate endoscopists from different hospital levels to generate standardized endoscopic reports.

CONFLICTS OF INTEREST

Guarantor of the article: Yan-qing Li, MD, PhD, and Xiu-li Zuo, MD, PhD.

Specific author contributions: Jun-yan Qu, MD and Zhen Li, MD, PhD, contributed equally to this work. Jun-yan Qu, MD, and Zhen Li, MD, PhD, contributed equally to this work. Conception and design: J.-y.Q, Z.L., Y.-q.L., and X.-l.Z. Analysis and interpretation of the data: J.-y.Q, Z.L., J.-r.S., M.-j.M., C.-q.X., A.-j.Z., C.-x.L., H.-p.Y., Y.-l.C., C.-c.L., L.-y. H., and L.L. Drafting of the article: J.-y.Q. and Z.L. Critical revision for important intellectual content: Z.L., Y.-q.L., and X.-l.Z. Final approval of the article: Y.-q.L. and X.-l.Z.

Financial support: This study was funded by the Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (NO.2019JZZY011007), the National Key R&D Program of China (No.2018YFB1307700), and clinical practical new technique development fund of Qilu Hospital (2019-13).

Potential competing interests: Xue-jun Shao, Yong-hang Lai, and Jian Feng are employed by Qingdao Medicon Digital Engineering Co, Ltd. The authors declare no conflict of interest.

Study Highlights

WHAT IS KNOWN

- ✓ GI endoscopy permits the early detection of digestive diseases.
- ✓ The conventional GI endoscopy report system was subjective and time consuming.
- ✓ The performance of AI in multitarget recognition of 10 major GI diseases remains unclear.

WHAT IS NEW HERE

- ✓ We developed and validated an ISRGS using a multicenter data set.
- ✓ The system could be a powerful detection tool for major GI lesions and their subcharacteristics.

TRANSLATIONAL IMPACT

- ✓ The system might facilitate endoscopists from different hospital levels to generate standardized endoscopic reports.

ACKNOWLEDGMENTS

We thank Xue-jun Shao, Yong-hang Lai, and Jian Feng for the technical support of artificial intelligence algorithm and Hao-wen Zhang and Iqtida Ahmed Mirza for English polishing, and Editage (www.editage.cn) for English language editing.

REFERENCES

1. Smith RA, Andrews KS, Brooks D, et al. Cancer screening in the United States, 2018: A review of current American cancer society guidelines and current issues in cancer screening. CA Cancer J Clin 2018;68:297–316.
2. Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin 2017; 67:93–9.
3. de Franchis R. Expanding consensus in portal hypertension: Report of the Baveno VI Consensus Workshop: Stratifying risk and individualizing care for portal hypertension. J Hepatol 2015;63:743–52.
4. Bisschops R, East JE, Hassan C, et al. Advanced imaging for detection and differentiation of colorectal neoplasia: European society of gastrointestinal endoscopy (ESGE) guideline–update 2019. Endoscopy 2019;51:1155–79.
5. Early Cancer Endoscopic Diagnosis and Treatment Cooperation Group, Digestive Department, Digestive Endoscopy Branch, Chinese Medical Association, et al. Chinese consensus: Screening, diagnosis and treatment of early esophageal squamous cell carcinoma and precancerous lesions (2015, Beijing) [in Chinese]. Chin J Pract Intern Med 2016;36:20–33.
6. Ono H, Yao K, Fujishiro M, et al. Guidelines for endoscopic submucosal dissection and endoscopic mucosal resection for early gastric cancer. Dig Endosc 2016;28:3–15.
7. Lee HL, Eun CS, Lee OY, et al. When do we miss synchronous gastric neoplasms with endoscopy? Gastrointest Endosc 2010;71:1159–65.
8. Kaminski MF, Wieszczy P, Rupinski M, et al. Increased rate of adenoma detection associates with reduced risk of colorectal cancer and death. Gastroenterology 2017;153:98–105.
9. Rees CJ, Thomas Gibson S, Rutter MD, et al. UK key performance indicators and quality assurance standards for colonoscopy. Gut 2016;65: 1923–9.
10. Miller AT, Sedlack RE. Competency in esophagogastroduodenoscopy: A validated tool for assessment and generalizable benchmarks for gastroenterology fellows. Gastrointest Endosc 2019;90:613–20.e1.
11. Han S, Obuch J, Keswani R, et al. A prospective multicenter study evaluating endoscopy competence among gastroenterology trainees in the era of the next accreditation system (NAS)–the EnCompAS study. Am J Gastroenterol 2018;113:S299.
12. Scaffidi MA, Grover SC, Carnahan H, et al. Impact of experience on self-assessment accuracy of clinical colonoscopy competence. Gastrointest Endosc 2018;87:827–36.e2.
13. Jamil LH, Naveed M, Agrawal D, et al. ASGE guideline on minimum staffing requirements for the performance of GI endoscopy. Gastrointest Endosc 2020;91:723–9.e17.
14. Hamashima C, Goto R. Potential capacity of endoscopic screening for gastric cancer in Japan. Cancer Sci 2017;108:101–7.
15. Horie Y, Yoshio T, Aoyama K, et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. Gastrointest Endosc 2019;89:25–32.
16. Byrne MF, Chapados N, Soudan F, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut 2019;68:94–100.
17. Luo H, Xu G, Li C, et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: A multicentre, case-control, diagnostic study. Lancet Oncol 2019;20:1645–54.
18. Ding Z, Shi H, Zhang H, et al. Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model. Gastroenterology 2019;157:1044–54.
19. Yao K. The endoscopic diagnosis of early gastric cancer. Ann Gastroenterol 2013;26:11–22.
20. Lee CT, Chang CY, Lee YC, et al. Narrow-band imaging with magnifying endoscopy for the screening of esophageal cancer in patients with primary head and neck cancers. Endoscopy 2010;42:613–9.
21. Hori K, Okada H, Kawahara Y, et al. Lugol-voiding lesions are an important risk factor for a second primary squamous cell carcinoma in patients with esosphageal cancer or head and neck cancer. Am J Gastroenterol 2011;106:858–66.
22. Nagami Y, Tominaga K, Machida H, et al. Usefulness of non-magnifying narrow-band imaging in screening of early esophageal squamous cell carcinoma: A prospective comparative study using propensity score matching. Am J Gastroenterol 2014;109:845–54.
23. Arantes V, Albuquerque W, Salles JM, et al. Effectiveness of unsedated transnasal endoscopy with white-light, flexible spectral imaging color enhancement, and lugol staining for esophageal cancer screening in high-risk patients. J Clin Gastroenterol 2013;47:314–21.
24. Chiu PWY, Uedo N, Singh R, et al. An Asian consensus on standards of diagnostic upper endoscopy for neoplasia. Gut 2019;68:186–97.
25. Bretthauer M, Aabakken L, Dekker E, et al. Requirements and standards facilitating quality improvement for reporting systems in gastrointestinal endoscopy: European society of gastrointestinal endoscopy (ESGE) position statement. Endoscopy 2016;48:291–4.
26. Su JR, Li Z, Shao XJ, et al. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: A prospective randomized controlled study. Gastrointest Endosc 2020;91:415–24.
27. Li J, Xu R, Liu M, et al. Lugol chromoendoscopy detects esophageal dysplasia with low levels of sensitivity in a high-risk region of China. Clin Gastroenterol Hepatol 2018;16:1585–92.
28. Ren W, Yu J, Zhang ZM, et al. Missed diagnosis of early gastric cancer or high-grade intraepithelial neoplasia. World J Gastroenterol 2013;19: 2092–6.
29. Zhao S, Wang S, Pan P, et al. Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: A systematic review and meta-analysis. Gastroenterology 2019;156:1661–76.e11.