Evaluation of the effects of an artificial intelligence system on endoscopy quality and preliminary testing of its performance in detecting early gastric cancer: a randomized controlled trial

Authors

Lianlian Wu[*,1,2,3], Xinqi He[*,1,2,3], Mei Liu[4], Huaping Xie[4], Ping An[1,2,3], Jun Zhang[1,2,3], Heng Zhang[5], Yaowei Ai[6], Qiaoyun Tong[7], Mingwen Guo[6], Manling Huang[5], Cunjin Ge[7], Zhi Yang[7], Jingping Yuan[8], Jun Liu[1,3], Wei Zhou[1,2,3], Xiaoda Jiang[1,2,3], Xu Huang[1,2,3], Ganggang Mu[1,2,3], Xinyue Wan[1,2,3], Yanxia Li[1,2,3], Hongguang Wang[9], Yonggui Wang[10], Hongfeng Zhang[11], Di Chen[1,2,3], Dexin Gong[1,2,3], Jing Wang[1,2,3], Li Huang[1,2,3], Jia Li[1,2,3], Liwen Yao[1,2,3], Yijie Zhu[1,2,3], Honggang Yu[1,2,3]

Institutions

1 Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, China
2 Key Laboratory of Hubei Province for Digestive System Disease, Renmin Hospital of Wuhan University, Wuhan, China
3 Hubei Provincial Clinical Research Center for Digestive Disease Minimally Invasive Incision, Renmin Hospital of Wuhan University, Wuhan, China
4 Department of Gastroenterology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
5 Department of Gastroenterology, Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
6 Department of Gastroenterology, The People's Hospital of China Three Gorges University/The First People's Hospital of Yichang, Yichang, China
7 Department of Gastroenterology, Yichang Central People's Hospital, China Three Gorges University, Yichang, China
8 Department of Pathology, Renmin Hospital of Wuhan University, Wuhan, China
9 Department of Gastroenterology, Jilin People's Hospital, Jilin, China
10 School of Geography and Information Engineering, China University of Geosciences, Wuhan, China
11 Department of Pathology, Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

submitted 21.9.2020
accepted after revision 11.1.2021
published online 11.1.2021

Bibliography
Endoscopy 2021; 53: 1199–1207
DOI 10.1055/a-1350-5583
ISSN 0013-726X

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Supplementary material
Supplementary material is available under https://doi.org/10.1055/a-1350-5583

Scan this QR-Code for the author commentary.

Corresponding author
Honggang Yu, MD, Department of Gastroenterology, Renmin Hospital of Wuhan University, 99 Zhangzhidong Road, Wuhan, China, 430060
yuhonggang@whu.edu.cn

ABSTRACT

Background Esophagogastroduodenoscopy (EGD) is a prerequisite for detecting upper gastrointestinal lesions especially early gastric cancer (EGC). An artificial intelligence system has been shown to monitor blind spots during EGD. In this study, we updated the system (ENDOANGEL), verified its effectiveness in improving endoscopy quality, and pretested its performance in detecting EGC in a multicenter randomized controlled trial.

Methods ENDOANGEL was developed using deep convolutional neural networks and deep reinforcement learning. Patients undergoing EGD in five hospitals were randomly assigned to the ENDOANGEL-assisted group or to a control

* These authors contribute equally to this work.

group without use of ENDOANGEL. The primary outcome was the number of blind spots. Secondary outcomes included performance of ENDOANGEL in predicting EGC in a clinical setting.

Results 1050 patients were randomized, and 498 and 504 patients in the ENDOANGEL and control groups, respectively, were analyzed. Compared with the control group, the ENDOANGEL group had fewer blind spots (mean 5.38 [standard deviation (SD) 4.32] vs. 9.82 [SD 4.98]; $P<0.001$) and longer inspection time (5.40 [SD 3.82] vs. 4.38 [SD 3.91] minutes; $P<0.001$). In the ENDOANGEL group, 196 gastric lesions with pathological results were identified. ENDOANGEL correctly predicted all three EGCs (one mucosal carcinoma and two high grade neoplasias) and two advanced gastric cancers, with a per-lesion accuracy of 84.7 %, sensitivity of 100 %, and specificity of 84.3 % for detecting gastric cancer.

Conclusions In this multicenter study, ENDOANGEL was an effective and robust system to improve the quality of EGD and has the potential to detect EGC in real time.

Introduction

Esophagogastroduodenoscopy (EGD) is widely used to examine upper gastrointestinal lesions [1, 2]. White-light imaging (WLI) endoscopy is a standard protocol for examining gastric lesions; however, the performance of endoscopists varies greatly, leading to a miss rate of 20 %–40 % for early gastric cancer (EGC) [3]. Endoscopy diagnosis is subjective, operator dependent, and varies widely with experience [4], reducing the detection rate of EGC and precursor lesions [5]. There is an urgent need to improve endoscopy quality and reliability.

To achieve such improvement, a large number of guidelines have been issued and consensus of expert opinions in specific areas has been reached [6]. Safety and quality indicators for EGD have been proposed by the American Society for Gastrointestinal Endoscopy and the American College of Gastroenterology [7]. The first evidence-based indicator of EGD performance was proposed by the European Society of Gastrointestinal Endoscopy (ESGE) in 2015 [1]. The standard procedure is to examine all parts of the stomach during EGD, with a recommended examination time of 7 minutes [8–10]. However, due to the lack of monitoring and available tools, adherence to protocols are often not very high [11]. A practical and workable approach should be established to implement guidelines for routine endoscopy.

In the past few years, deep learning has made remarkable progress in the field of medical image recognition [12]. Most studies are dedicated to the use of computer-aided diagnosis of lesions [13, 14]; however, whether deep convolutional neural networks (CNNs) can be used to monitor the quality of routine endoscopies has rarely been explored. In a previous study, our group developed a novel artificial intelligence (AI) system, named WISENSE, based on deep reinforcement learning (DRL) and CNN. WISENSE demonstrated the ability to monitor blind spots (gastric areas overlooked during EGD) and generate photodocumentation in real time during EGD [15, 16]. In the present study, we updated the WISENSE system by integrating a previously trained real-time EGC detection model [15], and named the updated system "ENDOANGEL." We then carried out a multicenter randomized controlled trial (RCT) to verify the ability of ENDOANGEL to improve EGD quality in five hospitals, and to describe its performance in detecting EGC in the clinical setting.

Methods

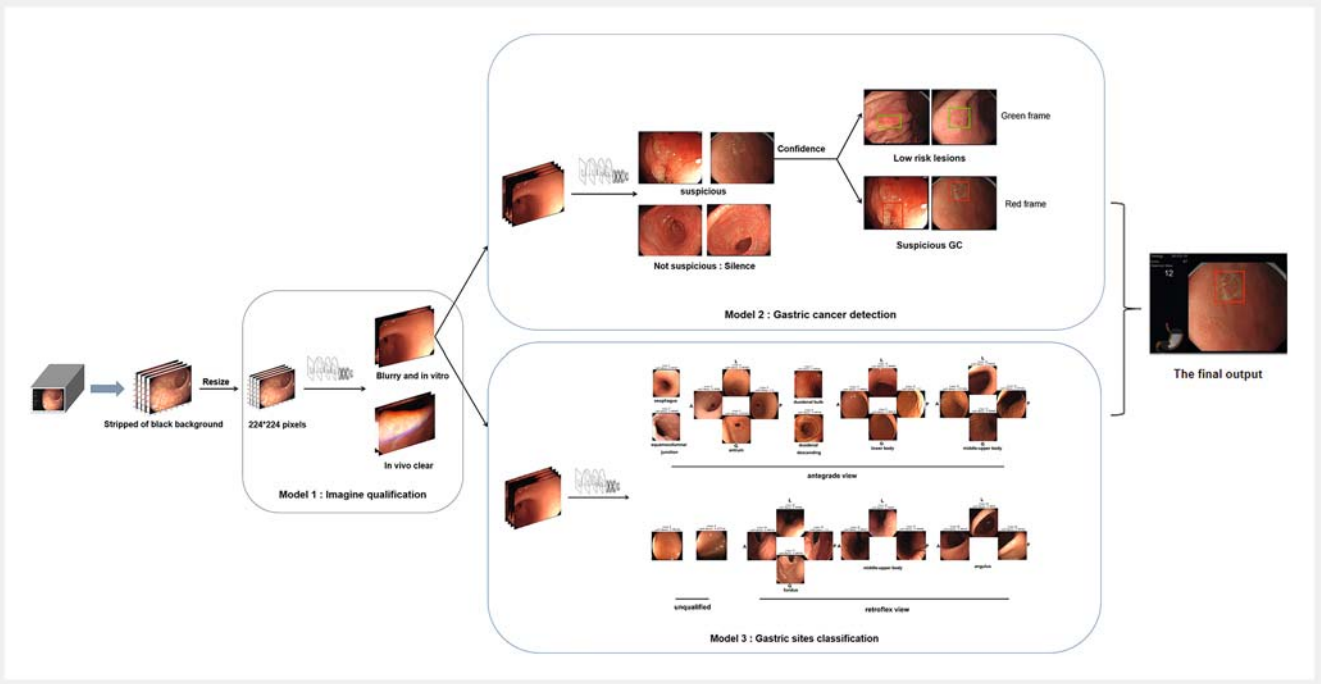
Development of the AI system

Three models – model 1 for image qualification, model 2 for gastric cancer prediction, and model 3 for gastric site classification – were involved in ENDOANGEL. Model 1 and 3 were trained as described in our previous single-center clinical trial [16], and model 2 was trained as described in our previous technical work [15]. Briefly, the VGG-16 CNN model was trained using transfer learning [17] with 12220 in vitro images, 25222 in vivo images, and 16760 unqualified images, which were filtered to retrieve only clear in vivo frames; the model achieved an accuracy of 97.6 % in 3000 still images (model 1). VGG-16 and ResNet-50 were respectively trained using transfer learning with 2204 EGC, 326 advanced gastric cancer, and 4791 noncancerous images, and achieved an accuracy of 92.5 % for predicting EGC in 200 still images when 3 VGG-16 and 2 Resnet-50 were combined (model 2). VGG-16 was trained using transfer learning with 34513 labeled EGD images of 26 different EGD sites, and DRL was trained using virtual EGD videos and 30 stored videos in order to achieve human logicality; VGG-16 combined with DRL achieved an accuracy of 90.0 % for predicting the gastric site in 107 real videos (model 3). Before images were fed to the CNN, they were first stripped of black borders and then resized to 224 × 224 pixels to suit the original dimensions of the CNN models. For the detection of EGC, a CNN algorithm was used. For the monitoring of blind spots, both CNN and DRL were implemented.

A few modifications were made to model 2 when it was integrated into the AI system, as described in the supplementary methods (see the online-only Supplementary material).

The three models were integrated, as illustrated in ▶ **Fig.1**, and frame-wise prediction was applied in a clinical setting using client–server interaction [15]. As tested in our previous work, the mean (standard deviation [SD]) total time to output of a prediction using all three models for each frame was 230 (SD 60) milliseconds. Therefore, ENDOANGEL was set to process EGD videos with 2 frames per second in real time.

The equipment used in this trial is described in the Supplementary material.



▶ **Fig. 1** Illustration diagram for integrating convolutional neural network (CNN) models into the ENDOANGEL system. Consecutive frames during esophagogastroduodenoscopy were first stripped of black borders and then resized to 224 × 224 pixels to suit the original dimensions of the CNN models. Then, the fitted frames were inputted into CNN1 for image qualification (Model 1), from which the blurry and in vitro images were discarded, and in vivo clear images were sent to Model 2 for gastric cancer (GC) detection and Model 3 for gastric site classification. The final output contains predictions of observed sites and a box localizing gastric cancer.

RCT trial design

This was a prospective, multicenter, single-blind, randomized, parallel-group study, approved by the Ethics Committee of Renmin Hospital of Wuhan University.

Patients

From October 2018 to January 2019, patients undergoing routine EGD examinations at the endoscopy centers of five tertiary hospitals were enrolled in the study. An introduction and details of the time period for patient enrollment in the five hospitals is presented in the Supplementary material. The RCT was approved by the institutional review boards of each participating hospital and performed according to the Declaration of Helsinki.

Inclusion criteria were: 1) age 18 years or above; 2) American Society of Anesthesiologists physical status score of 1, 2, or 3; 3) informed consent provided.

Exclusion criteria were: 1) patients with absolute contraindications to EGD examination; 2) history of previous gastric surgery; 3) pregnancy; 4) previous medical history of allergic reaction to anesthetics; 5) unsuitability for participation in the trial at the investigator's discretion. Withdrawal criteria were: 1) EGD surgery not completed due to esophageal stenosis, obstruction, large space-occupying lesions, or ulcers in the duodenal bulb; 2) premature termination of the EGD due to rapid changes in the patient's heart rate or respiratory rate.

The patient population was not limited to specific indications, as most patients with EGC are asymptomatic [18].

Before the trial, 14 enrolled endoscopists studied the ENDOANGEL user interface and the Japanese systematic screening protocol for the stomach [10]. The participating endoscopists at the five hospitals included six from Renmin Hospital of Wuhan University, two from Tongji Hospital, two from Central Hospital of Wuhan, two from Yichang Central People's Hospital, and two from the First People's Hospital of Yichang. The participating endoscopists had 3–5 years of EGD experience, had performed 2000–5000 EGD examinations, had diagnosed < 200 EGC cases, and had the ability to evaluate gastric lesions with magnifying image-enhanced endoscopy (M-IEE).

Interventions

Patients undergoing EGD examination were randomly assigned to a procedure with ENDOANGEL assistance or no assistance (control). The examination protocol consisted of WLI observation, M-IEE observation, and biopsy of suspicious lesions. In both groups, endoscopists first screened the upper gastrointestinal tract using WLI. A biopsy was taken if the endoscopist predicted that a lesion had a risk of gastric cancer. When endoscopists could not determine the risk of the lesion using WLI, M-IEE was used to make further observations and take targeted biopsies. In addition to the original video, four additional pieces of information were provided to the endoscopists in the ENDOANGEL group: 1) a virtual stomach model monitoring blind spots; 2) procedure time and duration; 3) red or green frames indicating cancerous and noncancerous lesions predicted by ENDOANGEL; 4) scoring and grading. The score was positively

correlated with the number of observed sites; scores of 80, 90, and 100 corresponded to "good," "excellent," and "perfect," respectively. No additional information was provided to the endoscopists in the control group. A working example of the ENDOANGEL system is shown in **Fig. 1s**, ▶ **Video 1**, and ▶ **Video 2**, and a representative image for ENDOANGEL detecting EGC is shown in **Fig. 2s**.

Outcomes

The main outcome of the study was the number of blind spots (out of 26 per patient) for both the ENDOANGEL and control groups. Blind spots were defined as the sites unobserved during EGD, indicated as transparent areas in the gastric icon, as shown in ▶ **Video 1** and ▶ **Video 2**.

The secondary outcomes were: 1) inspection duration; 2) the percentage of patients with missing observations (i. e. blind spots) at each site; 3) performance of ENDOANGEL in predicting EGC in a clinical setting.

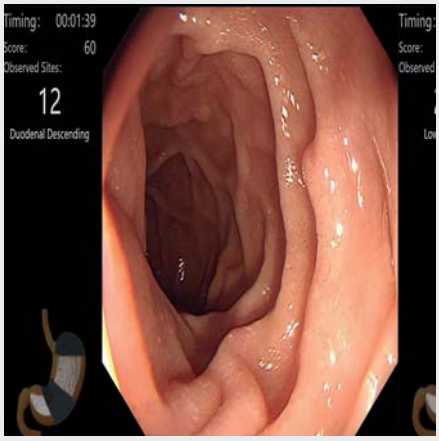
The number of blind spots, inspection duration, and the percentage of patients with blind spots at each site were analyzed in both groups, whereas the performance of ENDOANGEL in predicting gastric cancer in the clinical setting was analyzed only in the ENDOANGEL group.

The performance of ENDOANGEL in detecting gastric cancer was analyzed using accuracy, sensitivity, and specificity. Accuracy was calculated as the number of true predictions divided by the total number of lesions; sensitivity was calculated as the number of correctly predicted gastric cancers divided by the total number of gastric cancers; specificity was calculated as the number of correctly predicted noncancerous lesions divided by the total number of noncancerous lesions. Noncancerous lesions included adenoma, low grade neoplasia, intestinal metaplasia, atrophic gastritis, nonatrophic gastritis, benign ulcer, polyps, Xanthoma, etc.

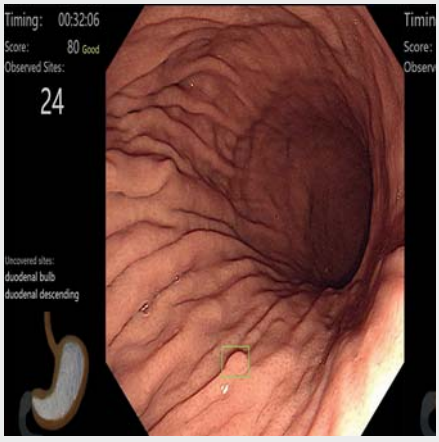
Two medical students reviewed the EGD data and recorded the start and end times of each EGD examination. Three experts with more than 10 years of EGD experience independently reviewed the EGD data from the trial patients and recorded the blind spots. A site was labeled as observed in a patient only when two or more experts reached an agreement. When the expert whose label was discarded had objections, the three experts would discuss the data together and reach a consensus. Endoscopists who performed the EGD examination did not participate in data evaluation.

Sample size

As the number of blind spots is a discrete variable, the sample size was calculated using the method of two-sample superiority tests. The mean number of blind spots in the control group and ENDOANGEL group were estimated as 10 and 5, respectively, with an overall SD of 4.3. With a power of 0.90, bilateral significance level of 0.05, and superiority margin of 4.2, 495 patients would be needed in each group. Assuming a dropout rate of 5 %, the target sample size for each group was 521.



Video 1 Representative video of the use of ENDOANGEL for monitoring blind spots and detecting noncancerous lesions. The system presented the covered gastric sites synchronized with the process of endoscopy to verify that the entire stomach was mapped. A cartoon gastric icon was set to be transparent before the examination. As soon as the scope was inserted into the stomach, the observed sites were colored in the corresponding part of the icon. Any transparent area indicated that the corresponding sites had not been observed (i. e. the blind spots). Meanwhile, ENDOANGEL successfully detected the gastric polyp and recognized it as a noncancerous lesion (in green box).
Online content viewable at:
https://doi.org/10.1055/a-1350-5583



Video 2 Representative video of the use of ENDOANGEL for detecting cancerous lesions. A pathologically confirmed early gastric cancer was shown in the video. ENDOANGEL successfully detected the lesion and recognized it as a suspicious cancerous lesion (in red box).
Online content viewable at:
https://doi.org/10.1055/a-1350-5583

Randomization and blinding

A computer-generated random numerical series was used to generate a random allocation sequence, with the ENDOANGEL group encoded as "0" and the control group as "1." Stratified randomization based on endoscopists was conducted in blocks of four in a 1:1 ratio. Endoscopists and statisticians were unblinded, whereas patients and all image data evaluations were performed blindly.

Statistical analysis

A chi-squared test was used to compare the ENDOANGEL and control groups in terms of baseline characteristics and the percentage of patients with blind spots at each site. The Mann–Whitney *U* test with a two-sided significance level of 0.05 was used to compare the other main and secondary outcomes between the two groups. The 95 % confidence intervals (CIs) with accuracy, sensitivity, and specificity were calculated using the method of Wilson procedure, with a correction for continuity. The receiver operating characteristic curve (ROC) was used to evaluate the performance of the CNN model for detecting EGC. The ROC curve was developed by plotting the sensitivity against the false-positive rate (i. e. 1-specificity) by varying prediction thresholds (**Fig. 3 s**). Statistical analysis was performed using StatsDirect version 3.1.20 (StatsDirect Ltd., Birkenhead, UK).

Results

Recruitment

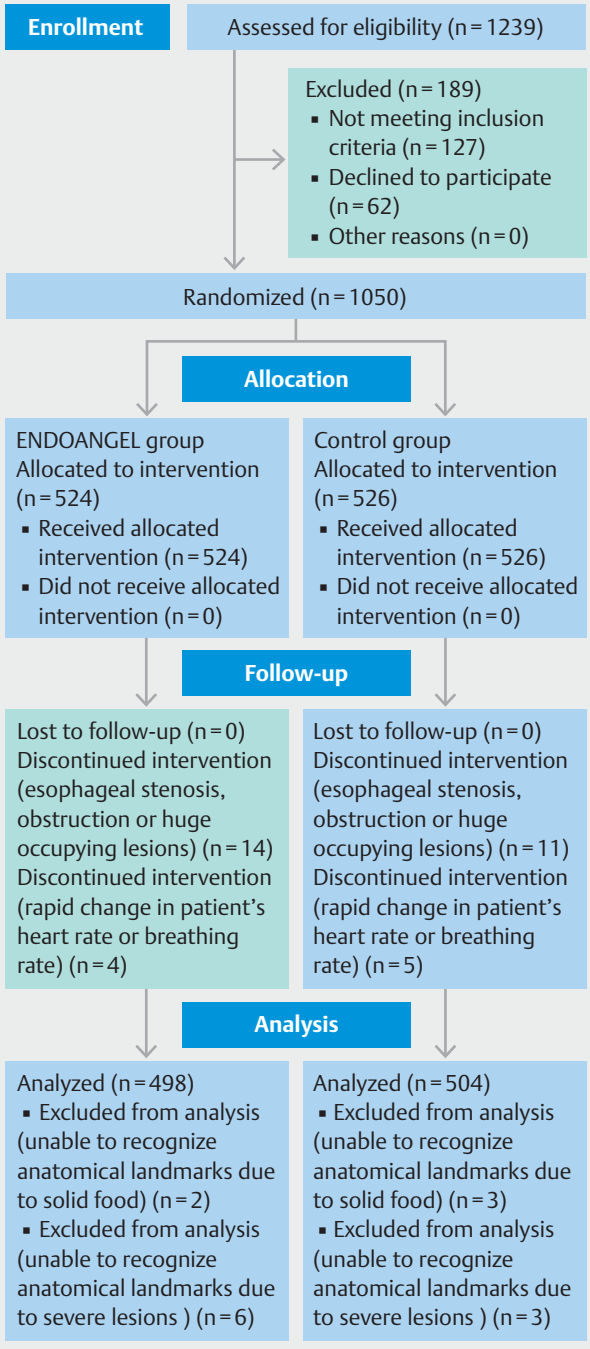
A total of 1239 patients were invited to participate in the trial, 189 of whom were excluded because they were ineligible (n = 127) or declined to participate (n = 62); therefore 1050 patients were recruited and randomized (▶ **Fig.2**). A total of 498 patients in the ENDOANGEL group and 504 in the control group were included in the final analysis of number of blind spots and other outcomes. Patient characteristics were comparable in both groups (▶ **Table 1**).

Blind spots and inspection duration

In the ENDOANGEL group, the mean number of blind spots was less than that in the control group (5.38 [SD 4.32] vs. 9.82 [SD 4.98]; $P<0.001$) (▶ **Table 2**). Mean inspection time of the EGD procedure was longer in the ENDOANGEL group than in the control group (5.40 [SD 3.82] minutes vs. 4.38 [SD 3.91] minutes; $P<0.001$) (▶ **Table 2**).

The median percentage of patients with blind spots at each site was 21.0 % (range 1.6 %–40.2 %) in the ENDOANGEL group and 38.9 % (range 0.8 %–68.3 %) in the control group. For 88.5 % of gastric sites (23/26), the percentage of patients in whom the site was overlooked was significantly lower in the ENDOANGEL group than in the control group (▶ **Table 3**).

The number of blind spots, with or without AI, were compared among the 14 endoscopists (**Fig.4 s**, **Table 1 s**). With the assistance of ENDOANGEL, the number of blind spots of 11 endoscopists significantly decreased, while that of the other 3 endoscopists had no significant change.



▶ **Fig. 2** Trial flow diagram.

Gastric cancer detection

Lesion characteristics in the ENDOANGEL group

In the 498 patients in the ENDOANGEL group, 819 lesions were reported by endoscopists. Of these lesions, 210 (25.6 %) had biopsy samples taken (196 gastric, 12 esophageal, and 2 duodenal lesions). The remaining 609 lesions, without biopsies, included 437 gastric, 90 esophageal, and 82 duodenal lesions. Lesion characteristics are described in **Table 2 s**. The number of images used per patient was 600 (interquartile range [IQR]

▶ **Table 1** Baseline characteristics.

Characteristics	ENDOANGEL (n = 498)	Control (n = 504)
Age, mean (SD), years	51.5 (13.2)	51.6 (13.1)
Female, n (%)	273 (54.8)	277 (55.0)
Indications for EGD, n (%)		
▪ Abdominal discomfort	359 (72.1)	366 (72.6)
▪ Acid reflux	36 (7.2)	33 (6.5)
▪ Anemia	2 (0.4)	5 (1.0)
▪ Belching	4 (0.8)	4 (0.8)
▪ Bowel habit change	6 (1.2)	2 (0.4)
▪ Constipation	3 (0.6)	3 (0.6)
▪ Diarrhea	8 (1.6)	11 (2.2)
▪ Dyspepsia	13 (2.6)	11 (2.2)
▪ Dysphagia	3 (0.6)	4 (0.8)
▪ Emaciation	3 (0.6)	4 (0.8)
▪ Health examination	19 (3.8)	18 (3.6)
▪ Poor appetites	2 (0.4)	2 (0.4)
▪ Suspected GI bleeding	18 (3.6)	20 (4.0)
▪ Suspected malignancy	14 (2.8)	17 (3.4)
▪ Vomiting	8 (1.6)	4 (0.8)
Recruitment, n (%)		
▪ Inpatient	165 (33.1)	175 (34.7)
▪ Outpatient	333 (66.9)	329 (65.3)

EGD, esophagogastroduodenoscopy; GI, gastrointestinal.

369–710) in the ENDOANGEL group and 485 (IQR 247–626) in the control group.

Real-time performance of ENDOANGEL in predicting gastric cancer in clinical practice

The two advanced gastric cancers and three EGCs confirmed by pathology in the ENDOANGEL group were positively predicted by ENDOANGEL. Among 302 692 EGD frames from 498 patients in the ENDOANGEL group, 2107 (0.7 %) red boxes indicating suspicious gastric cancer were included in the ENDOANGEL outputs. Of these, 357 (16.9 %) were diagnosed by endoscopists to have lesions, and the remaining 1750 red boxes contained "noise," including reflections, foam, mucus, and folds, as summarized in **Table 3 s**. For 196 gastric lesions with pathological results, ENDOANGEL correctly predicted all 5 gastric cancers (2 advanced gastric cancer, 1 mucosal carcinoma, and 2 high grade neoplasia), with a per-lesion accuracy of 84.7 % (95 %CI 78.7 %–89.3 %), sensitivity of 100 % (95 %CI 46.3 %–100 %), and specificity of 84.3 % (95 %CI 78.2 %–89.0 %). For 437 gastric lesions with no pathological results, 31 (7.1 %) were positively predicted by ENDOANGEL, with the highest positive prediction rates shown for hemorrhagic gastritis (16.7 % [1/6]), protruding lesions (15.8 % [3/19]), and erosive gastritis (11.7 % [19/163]).

Discussion

Gastric cancer is the third leading cause of cancer death from a global perspective [19]. Early detection is the key strategy to improve patient survival. However, the quality of endoscopy varies significantly, impairing the health outcome of patients. Technically, complete observation is an essential prerequisite for detecting EGC; however, although protocols for mapping the entire stomach have been widely proposed, they are often not followed closely in clinical practice. Cognitively, EGC lesions are difficult to recognize because the mucosal changes are often very subtle, requiring endoscopists to have thorough knowledge and extensive experience [4, 7]. In the current study, we developed ENDOANGEL, a real-time AI assistance system for the detection of EGC, with no blind spots, to specifically address these two problems. In this multicenter RCT of blind spot monitoring, we validated effectiveness and robustness of ENDOANGEL in improving EGD quality; in addition, we prospectively evaluated the performance and feasibility of ENDOANGEL for the detection of EGC in clinical practice.

Gastric cancer may occur in every part of the gastric cavity [20]. Endoscopist competence is an essential prerequisite for the detection of EGC lesions during EGD. In our previous work, we developed an AI system to classify different gastric sites and monitor blind spots in real time during EGD, and verified the effectiveness of the system in improving EGD quality in a single-center RCT [15, 16]. Results from this single-center study showed that the number of blind spots dropped from 5.84 to 1.52 with the assistance of AI. In the current multicenter RCT, we further verified the effect of improving EGD quality in five different hospitals, and the number of blind spots dropped from 9.82 to 5.38 with the assistance of ENDOANGEL. The findings of the two studies are consistent; however, in the present study, the number of blind spots was higher in both the ENDOANGEL and control groups compared with that in the previous single-center study, possibly as a result of variability in the operation quality across the hospitals. The effect of AI on endoscopist practice may be influenced by endoscopists' experience with AI systems and their personal views and acceptance of AI technology, according to a previous report [21]. In order to avoid possible center effects, several measures were implemented in the present study. First, the five hospitals included were all tertiary hospitals and the participating endoscopists were senior endoscopists with an EGD experience of 3–5 years and EGD volumes of 2000–5000 examinations. Second, to unify the endoscopic observation procedures the 14 participating endoscopists were trained to use the Japanese systematic screening protocol for the stomach before the trial. More importantly, results from each hospital, including the number of blind spots, were evaluated and analyzed by the same data analysis team; for cases in which the endoscopic results were inconsistent with the pathological results, data were reviewed

▶ **Table 2** Primary and secondary outcomes for all patients compared with results from our previous single-center trial.

End point	Mean (SD)		*P* value	Mean (SD)		*P* value
	ENDOANGEL (n = 498)	Control (n = 504)		WISENSE (n = 153)	Control (n = 150)	
Primary end point						
▪ No. of blind spots*	5.38 (4.32)	9.82 (4.98)	< 0.001	1.52 (1.79)	5.84 (3.73)	< 0.001
Secondary end point						
▪ Inspection time, minutes	5.40 (3.82)	4.38(3.91)	< 0.001	5.03 (2.95)	4.24 (3.82)	< 0.001

SD, standard deviation.

* The number of blind spots per patient out of a total of 26 gastric sites; Results for WISENSE were cited from our previous single-center clinical trial [16]. It should be noted that the primary outcome in the previous study is "blind spot rate," and the number of blind spots shown in the table (1.52 and 5.84) were converted from the rate of blind spots (5.86 % and 22.46 %) by multiplying by 26.

by a single expert pathologist, to reduce differences within and between centers.

Recently, several studies have tried to use deep learning for EGC recognition. Hirasawa et al. developed a CNN to detect gastric cancer, which achieved a sensitivity of 92.2 % [14]. Li et al. achieved a sensitivity of 91.2 % for detection of EGC in 341 still images [22]. Our group also developed a CNN model for detecting EGC, which achieved a sensitivity of 94.0 % in 200 still images [15]. However, the images chosen for testing in the previous studies were retrospectively selected, and the types of noncancerous lesions were limited. In the real world, plenty of lesions are difficult to distinguish from EGC, such as erosive gastritis and ulcers, and such lesions were uncommon in the testing datasets in the previous work. Such selection of lesions may lead to a bias in accuracy in favor of CNN models. In addition, there is a mass of "noise" during endoscopy in real clinical setting, such as reflections, blurring, and foam, whereas most retrospective images are of good quality. Therefore, we prospectively applied our previously trained EGC detection model in a multicenter clinical trial, evaluated its performance in complex clinical environments, and provided suggestions for further work in the development of EGC detection models.

Our results revealed two prominent problems when applying EGC detection models to clinical practice. First, "noise" greatly impacts the accuracy of the model and is bothersome to endoscopists. Images showing "noise" from consecutive frames in videos could be collected to train the model to recognize and filter out "noise." Methods including localization [23] and segmentation [24] could be explored to solve this problem by targeting and shielding the image "noise." Second, some endoscopists argue that it is almost impossible to accurately predict EGC in white-light view because other lesions such as erosive gastritis and ulcers share similar characteristics with EGC. The same point is also presented in the guidelines from ESGE [25]. Our results showed that a small proportion of benign lesions such as erosive gastritis, ulcers, and polyps were incorrectly diagnosed as gastric cancer, and they are difficult to distinguish from EGC even for experienced endoscopists. The quantity and diversity of training datasets could be further increased to improve the performance of the CNN model in order to extend the limits of human visualization and interpretation. In addition, we may change our minds and adjust the aims of the AI model from detection of EGC to recognition of abnormal lesions in WLI that need further observation with IEE. Further studies should be conducted to explore the supposed solutions and to improve the EGC detection model.

In the past few years, the performance of CNN models has been generally improved by increasing the depth and fitting parameters [24, 26]. In 2016, He et al. proposed the concept of residuals, and it was proved to be easier for optimization and achieved better performance with fewer parameters [27]. Nowadays, deeper models and smaller kernels are preferred over single layer and larger kernels [28]. Liu et al. elaborately compared different CNN models, with or without transfer learning, on classifying EGC and gastritis, and found that ResNet-50 achieved a top accuracy of 95 % when using transfer learning [29]. In the present study, we trained both ResNet-50 and VGG-16 using transfer learning for predicting EGC, and their combined results achieved an accuracy of 92.5 % in still images. The results of two types of CNNs were combined to reduce the rate of miss-selection of a single classifier [30]; however, in clinical practice, although all EGC lesions were successfully diagnosed, the false-positive rate increased. Some scholars have explored 3D-CNNs [31], segmentation, and long short-term memory network with CNN to improve prediction results [23]. These experiences are valuable for further research.

There are some limitations to our study. First, we only conducted a feasibility analysis on real-time detection of gastric cancer based on deep learning in a clinical setting. Whether the AI system can achieve a good performance in gastric cancer detection and help improve the detection rate of EGC remains to be investigated in larger multicenter studies. Second, the enrolled patients were not followed up for a long time, and this may lead to false-negative lesions missed by endoscopists, and the diagnostic ability of the endoscopists may have an impact

▶ **Table 3** The median percentage of patients with blind spots at each site compared with results from our previous single-center trial.

Overlooked sites	ENDOANGEL (n = 498)	Control (n = 504)	*P* value	WISENSE (n = 153)	Control (n = 150)
Esophagus	1.6	0.8	0.237	0	0
Squamocolumnar junction	6.2	9.3	0.067	0	1.33
Antrum (G)	9.0	14.1	0.012	0	3.33
Antrum (P)	21.1	39.3	<0.001	2.61	10.00
Antrum (A)	23.3	37.5	<0.001	2.61	6.67
Antrum (L)	16.1	31.7	<0.001	3.92	9.33
Duodenal bulb	4.8	7.1	0.121	0.65	4.00
Duodenal descending	1.6	5.4	0.001	0	6.00
Lower body (G)	8.2	21.6	<0.001	2.61	17.33
Lower body (P)	26.9	56.2	<0.001	13.07	29.33
Lower body (A)	20.9	43.3	<0.001	7.19	18.67
Lower body (L)	18.7	43.5	<0.001	5.23	30.00
Middle-upper body (F, G)	14.1	20.4	0.008	2.61	5.33
Middle-upper body (F, P)	35.3	60.7	<0.001	13.07	34.67
Middle-upper body (F, A)	40.2	68.3	<0.001	13.07	42.67
Middle-upper body (F, L)	38.8	64.3	<0.001	8.50	56.00
Fundus (G)	5.8	12.1	0.001	2.61	8.67
Fundus (P)	17.1	35.7	<0.001	8.50	21.33
Fundus (A)	21.3	38.5	<0.001	14.38	17.33
Fundus (L)	36.3	63.1	<0.001	18.95	40.67
Middle-upper body (R, P)	34.7	58.7	<0.001	6.54	17.33
Middle-upper body (R, A)	32.1	57.7	<0.001	19.61	40.67
Middle-upper body (R, L)	33.5	50.0	<0.001	13.73	24.00
Angulus (P)	33.1	63.7	<0.001	27.45	64.00
Angulus (A)	27.5	56.7	<0.001	12.42	53.33
Angulus (L)	9.8	22.4	<0.001	3.27	19.33

A, anterior wall; G, greater curvature; F, forward view; L, lesser curvature; P, posterior wall; R, retroflex view.

on the evaluation of the ENDOANGEL performance. To avoid this bias, further study in which all patients are followed up or biopsied should be conducted in order to evaluate the precision of ENDOANGEL in detecting gastric cancer in a clinical setting. Third, patients and all image data evaluations were performed blindly in this trial, whereas statisticians were not blinded. Unblinded statisticians may induce potential bias in analysis; more attention should be paid to this issue in our future research. Fourth, in addition to IEE, chromoendoscopy is also one of the major tools used for tumor detection and characterization. In our previous work, a deep learning method was developed to delineate EGC margin under chromoendoscopy [32]; use of AI to detect and diagnose EGC under chromoendoscopy is still a valuable direction that could be tried in the future.

In conclusion, ENDOANGEL, a system for improving endoscopy quality based on deep learning, achieved real-time monitoring of endoscopic blind spots, timing, and EGC detection during EGD. ENDOANGEL greatly improved the quality of EGD in this multicenter study, and showed potential for detecting EGC in real clinical settings.

Acknowledgments

We thank our endoscopists and machine-learning engineers for their hard work. We express gratitude to all patients and hospital staff for support of our trial.

Clinical trial

Trial Registration: Chinese Clinical Trial Registry | Registration number (trial ID): ChiCTR1800018403 | Type of study: Randomized, Multi-Center Study

Funding

Project of Hubei Provincial Clinical Research Center for Digestive Disease Minimally Invasive Incision2018BCC337
Hubei Province Major Science and Technology Innovation Project2018–916–000–008

Competing interests

The authors declare that they have no conflict of interest.

References

[1] Bisschops R, Areia M, Coron E et al. Performance measures for upper gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. Endoscopy 2016; 48: 843–864

[2] Pasechnikov V, Chukov S, Fedorov E et al. Gastric cancer: prevention, screening and early diagnosis. World J Gastroenterol 2014; 20: 13842–13862

[3] Kaise M. Advanced endoscopic imaging for early gastric cancer. Best Pract Res Clin Gastroenterol 2015; 29: 575–587

[4] Axon A. Is diagnostic and therapeutic endoscopy currently appropriate? suggestions for improvement. Best Pract Res Clin Gastroenterol 2008; 22: 959–970

[5] Gado AS, Ebeid BA, Axon AT. Quality assurance in gastrointestinal endoscopy: an Egyptian experience. Arab J Gastroenterol 2016; 17: 153–158

[6] Malheiro R, de Monteiro-Soares M, Hassan C et al. Methodological quality of guidelines in gastroenterology. Endoscopy 2014; 46: 513–525

[7] Rizk MK, Sawhney MS, Cohen J et al. Quality indicators common to all GI endoscopic procedures. Gastrointest Endosc 2015; 81: 3–16

[8] Teh JL, Tan JR, Lau LJ et al. Longer examination time improves detection of gastric cancer during diagnostic upper gastrointestinal endoscopy. Clin Gastroenterol Hepatol 2015; 13: 480–487

[9] Ito Y, Blackstone MO. The endoscopic diagnosis of early gastric cancer. Gastrointest Endosc 1979; 25: 96–101

[10] Yao K. The endoscopic diagnosis of early gastric cancer. Ann Gastroenterol 2013; 26: 11–22

[11] Rutter MD, Rees CJ. Quality in gastrointestinal endoscopy. Endoscopy 2014; 46: 526–528

[12] Litjens G, Kooi T, Bejnordi BE et al. A survey on deep learning in medical image analysis. Med Image Anal 2017; 42: 60–88

[13] Urban G, Tripathi P, Alkayali T et al. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 2018; 155: 1069–1078

[14] Hirasawa T, Aoyama K, Tanimoto T et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer 2018; 21: 653–660

[15] Wu L, Zhou W, Wan X et al. A deep neural network improves endoscopic detection of early gastric cancer without blind spots. Endoscopy 2019; 51: 522–531

[16] Wu L, Zhang J, Zhou W et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut 2019; 68: 2161–2169

[17] Shao L, Zhu F, Li X. Transfer learning for visual categorization: a survey. IEEE Trans Neural Netw Learn Syst 2015; 26: 1019–1034

[18] Leung WK, Wu M-s, Kakugawa Y et al. Screening for gastric cancer in Asia: current evidence and practice. Lancet Oncol 2008; 9: 279–287

[19] Karimi P, Islami F, Anandasabapathy S et al. Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention. Cancer Epidemiol Biomarkers Prev 2014; 23: 700–713

[20] Huang Q, Shi J, Sun Q et al. Clinicopathological characterisation of small (2 cm or less) proximal and distal gastric carcinomas in a Chinese population. Pathology 2015; 47: 526–532

[21] Jin EH, Lee D, Bae JH et al. Improved accuracy in optical diagnosis of colorectal polyps using convolutional neural networks with visual explanations. Gastroenterology 2020; 158: 2169–2179

[22] Li L, Chen Y, Shen Z et al. Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow band imaging. Gastric Cancer 2020; 23: 126–132

[23] Tajbakhsh N, Gurudu SR, Liang J. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. Proceedings of the 12th International Symposium on Biomedical Imaging (ISBI); 2015 April 16–19; Brooklyn, NY, USA. IEEE 2015; July; 79–83.

[24] Hazirbas C, Ma L, Domokos C et al. Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture. Asian conference on computer vision. Cham: Springer2016: 213–228

[25] Dinis-Ribeiro M, Areia M, de Vries AC et al. Management of precancerous conditions and lesions in the stomach (MAPS): guideline from the European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter Study Group (EHSG), European Society of Pathology (ESP), and the Sociedade Portuguesa de Endoscopia Digestiva (SPED). Endoscopy 2012; 44: 74–94

[26] Widya AR, Monno Y, Okutomi M et al. Whole stomach 3D reconstruction and frame localization from monocular endoscope video. IEEE J Transl Eng Health Med 2019; 7: 1–10

[27] He K, Zhang X, Ren S et al. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 June 27–30; Las Vegas, NV, USA. IEEE 2016: 770–778.

[28] Du W, Rao N, Liu D et al. Review on the applications of deep learning in the analysis of gastrointestinal endoscopy images. IEEE Access 2019; 7: 142053–142069

[29] Liu X, Wang C, Bai J et al. Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images. Neurocomputing 2020; 392: 253–267

[30] Zhou Z. Machine learning (in Chinese) [M]. Beijing, China: Tsinghua University Press; 2016

[31] Lequan Y, Hao C, Qi D et al. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE J Biomed Health Inform 2017; 21: 65–75

[32] An P, Yang D, Wang J et al. A deep learning method for delineating early gastric cancer resection margin under chromoendoscopy and white light endoscopy. Gastric Cancer 2020; 23: 884–892