Artificial intelligence versus expert endoscopists for diagnosis of gastric cancer in patients who have undergone upper gastrointestinal endoscopy

OPEN ACCESS

Authors

Ryota Niikura[1,2], Tomonori Aoki[1], Satoki Shichijo[3], Atsuo Yamada[1] ⬤, Takuya Kawahara[4], Yusuke Kato[5], Yoshihiro Hirata[6], Yoku Hayakawa[1], Nobumi Suzuki[1], Masanori Ochi[1], Toshiaki Hirasawa[7], Tomohiro Tada[5,8,9], Takashi Kawai[2], Kazuhiko Koike[1]

Institutions

1   Department of Gastroenterology, Graduate School of Medicine, The University of Tokyo, Japan
2   Gastroenterological Endoscopy, Tokyo Medical University, Tokyo, Japan
3   Department of Gastrointestinal Oncology, Osaka International Cancer Institute, Osaka, Japan
4   Clinical Research Promotion Center, The University of Tokyo Hospital, Tokyo, Japan
5   AI Medical Service Inc., Tokyo, Japan
6   Division of Advanced Genome Medicine, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan
7   Department of Gastroenterology, Cancer Institute Hospital Ariake, Japanese Foundation for Cancer Research, Tokyo, Japan
8   Department of Surgical Oncology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
9   Tada Tomohiro Institute of Gastroenterology and Proctology, Saitama, Japan

🌐 Supplementary material
Supplementary material is available under
https://doi.org/10.1055/a-1660-6500

Corresponding author

Ryota Niikura, MD PhD, Gastroenterological Endoscopy, Tokyo Medical University, 6-7-1 Nishishinjuku, Shinjuku-ku, Tokyo 1600023, Japan
niikura-dky@umin.ac.jp

ABSTRACT

Aims To compare endoscopy gastric cancer images diagnosis rate between artificial intelligence (AI) and expert endoscopists.

Patients and methods We used the retrospective data of 500 patients, including 100 with gastric cancer, matched 1:1 to diagnosis by AI or expert endoscopists. We retrospectively evaluated the noninferiority (prespecified margin 5 %) of the per-patient rate of gastric cancer diagnosis by AI and compared the per-image rate of gastric cancer diagnosis.

Results Gastric cancer was diagnosed in 49 of 49 patients (100 %) in the AI group and 48 of 51 patients (94.12 %) in the expert endoscopist group (difference 5.88, 95 % confidence interval: −0.58 to 12.3). The per-image rate of gastric cancer diagnosis was higher in the AI group (99.87 %, 747/748 images) than in the expert endoscopist group (88.17 %, 693/786 images) (difference 11.7 %).

Conclusions Noninferiority of the rate of gastric cancer diagnosis by AI was demonstrated but superiority was not demonstrated.

Introduction

Upper gastrointestinal endoscopy is the standard procedure for diagnosis of gastric cancer. However, gastric cancer may be diagnosed within a few years after endoscopy because of missed lesions. Artificial intelligence (AI)-aided methods are needed to reduce the rate of missed lesions by automatic detection of gastric cancer, which could reduce the mortality rate.

AI based on deep learning shows promise for gastric cancer surveillance. Use of convolutional neural networks (CNNs) for deep learning enables extraction of specific features from endoscopic images and endoscopic diagnosis. Twelve previous studies, including ours [1], have investigated the diagnosis of gastric cancer lesions using upper gastrointestinal endoscopy images [2–11]. The results were heterogeneous, but most models reached a sensitivity of over 80 %. However, these studies had technical limitations, including problems with patient-level comparison of the efficacy of gastric cancer diagnosis by AI and by expert endoscopists. In addition, to evaluate gastric cancer diagnosis it is important to reduce bias and the influence of confounding factors. For these reasons, we conducted a retrospective matching analysis to evaluate noninferiority of the detection rate of gastric cancer by AI compared with that of expert endoscopists. A STROBE checklist statement for items that should be included in reports of observational studies has been completed for this study (Table 1 s in the online-only supplementary material).

Methods

Patients

We retrospectively selected patients aged 20 years or over who had previously undergone upper gastrointestinal endoscopy at the University of Tokyo Hospital during 2018. All upper gastrointestinal endoscopies were performed using an electronic video endoscope (Olympus Medical Systems, Tokyo, Japan). Indications for endoscopy were gastric cancer surveillance or gastroesophageal symptoms. Biopsy specimens were obtained from gastric cancer lesions. Histological diagnosis of gastric cancer was performed and confirmed by experienced pathologists. The trial was approved by the institutional review board of the University of Tokyo Hospital. The study protocol and statistical analysis plan were published before initiation of the study.

Preparation of the endoscopic image dataset and AI algorithm

We collected 23 892 white-light upper gastrointestinal endoscopy images of 500 patients, including 985 invasive gastric cancer images from 51 patients and 549 early gastric cancer images from 49 patients confirmed histologically. Early gastric cancer was defined as T1a and invasive gastric cancer as T1b–T4 (Union for International Cancer Control tumor–node–metastasis classification, v. 8).

The images were collected and prepared in July 2019. The investigators (R.N. and T.A.) annotated gastric cancer lesions with their coordinates (X, Y) in the images; gold-standard bounding boxes were generated, and data concealment was carried out. The AI algorithm method termed the Single Shot MultiBox Detector was used [1].

Trial design and diagnosis

Patients were matched (1:1) to diagnosis by AI or expert endoscopists using a computer-based matching system. Stratified matching of early and invasive gastric cancer and *Helicobacter pylori* status was performed in accordance with the allocation sequence generated by the trial statistician at the University of Tokyo. *H. pylori* status was defined as positive, negative, or eradicated, based on the most recent serological, urea breath test, or stool antigen test results.

After matching, endoscopic image diagnosis was performed by both AI and expert endoscopists. The optimal diagnostic cut-off for AI diagnosis was taken from a prior report [1]. The AI reviewed endoscopy images and reported those in which gastric cancer was detected, together with the coordinates (X, Y) of the lesions. The expert endoscopists, two physicians with experience of more than 20 000 endoscopies, reviewed the endoscopy images of each patient for 5 minutes and reported endoscopic images in which gastric cancer was detected; they manually annotated the coordinates (X, Y) of the lesions in those images.

Outcomes

The main outcome was per-patient diagnosis of gastric cancer. Detection of gastric cancer by AI and expert endoscopists on even one gastric cancer endoscopic image was defined as diagnosis of gastric cancer. The definition of accuracy was the presence of overlap between the AI-drawn bounding boxes with a probability score threshold of 0.01 or greater, expert endoscopist-drawn bounding boxes, and the gold-standard boxes in gastric cancer endoscopic images. If the AI drew multiple bounding boxes in the same gastric cancer lesion, we used the bounding box with the highest probability score.

Other outcomes were per-patient diagnosis of invasive gastric cancer, per-patient diagnosis of early gastric cancer, per-image diagnosis of gastric cancer, and intersection over union (IOU) of gastric cancer. Per-image diagnosis of gastric cancer was evaluated as the number of images analyzed for diagnosis of gastric cancer. IOU was defined as the amount of overlap between the area of the predicted and the gold-standard bounding boxes; it ranged from 0 to 1 (see online-only supplementary material, Fig.1 s).

Statistical analysis

Data regarding the per-patient rate of gastric cancer diagnosis, per-patient rate of invasive gastric cancer diagnosis, per-patient rate of early gastric cancer diagnosis, and per-image rate of gastric cancer diagnosis were compared by χ^2 test and risk difference assessment. IOU was compared by *t*-test and risk difference assessment. Analyses were performed using SAS software v. 9.4 (SAS Institute, Cary, North Carolina, USA).

▶ **Table 1** Baseline patient characteristics (n = 500).

Variable	AI diagnosis, n = 249	Expert endoscopist diagnosis, n = 251	P value
Age, mean ± SD, years	72.2 ± 9.54	72.0 ± 9.55	0.629
Sex, male	137 (55.02)[1]	136 (54.18)	0.851
Endoscopic atrophy[2]			
▪ No atrophy	88 (35.34)	87 (34.66)	0.873
▪ C-1	7 (2.81)	6 (2.39)	0.768
▪ C-2	29 (11.65)	17 (6.77)	0.059
▪ C-3	22 (8.84)	29 (11.55)	0.315
▪ O-1	30 (12.05)	31 (12.35)	0.918
▪ O-2	38 (15.26)	45 (17.93)	0.423
▪ O-3	36 (14.35)	35 (14.05)	0.927
H. pylori status[3]			
▪ Negative	123 (49.40)	123 (49.00)	0.982
▪ Positive	13 (4.82)	13 (5.18)	
▪ Eradicated	114 (45.78)	115 (45.82)	
Number of patients with gastric cancer	49 (19.68)	51 (20.32)	0.858
Early gastric cancer	27 (10.84)	26 (10.36)	0.860
Invasive gastric cancer	22 (8.84)	25 (9.96)	0.667
Number of gastric cancer images/nongastric cancer images	748 / 11 185 (6.27)	786 / 11 173 (6.57)	0.338

Abbreviations: AI, artificial intelligence; SD, standard deviation.
[1] Figures given in parentheses are percentages.
[2] Endoscopic atrophy was evaluated according to the Kimura–Takemoto classification, which considers no atrophy to grade C3 atrophy as closed type and grades O1 to O3 as open type; no atrophy was the mildest and O3 was the most severe. Closed type was milder than open type.
[3] *H. pylori* status was defined as: negative: *H. pylori* antibody, urea breath test (UBT), or *H. pylori* stool antigen test negative; positive: *H. pylori* antibody, UBT, or *H. pylori* stool antigen test positive; or eradicated: successful eradication confirmed by UBT or *H. pylori* stool antigen test after eradication therapy.
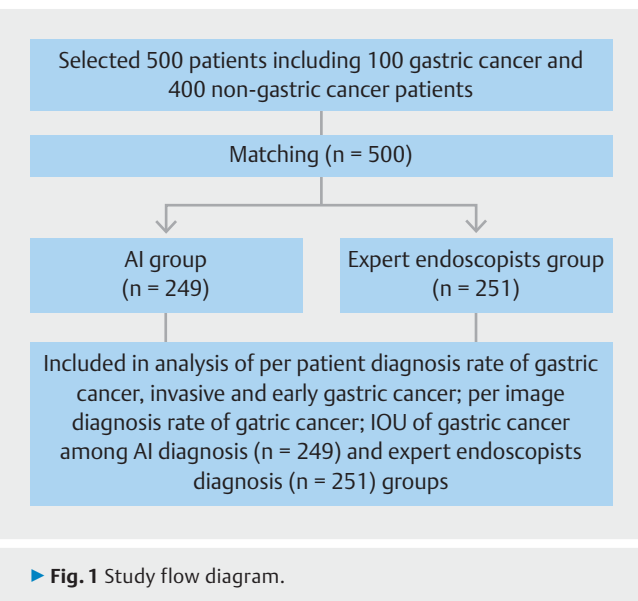
Results

Baseline characteristics

Of the 500 patients who underwent a matching analysis, 249 were allocated to the AI diagnosis group and 251 to the expert endoscopist diagnosis group (▶ **Fig.1**). Patient demographics were similar between the groups (▶ **Table 1**).

Outcomes

Gastric cancer was diagnosed in 49 of 49 patients (100 %) in the AI diagnosis group and 48 of 51 (94.12 %) in the expert endoscopist diagnosis group (difference 5.88, 95 % confidence interval [CI]: −0.58 to 12.3) (▶ **Table 2**). Invasive gastric cancer was diagnosed in 22 of 22 patients (100 %) in the AI diagnosis group and 25 of 25 patients (100 %) in the expert endoscopist diagnosis group. Early gastric cancer was diagnosed in 27 of 27 patients (100 %) in the AI diagnosis group and 23 of 26 patients (88.46 %) in the expert endoscopist diagnosis group (difference 11.54, 95 %CI −0.74 to 23.82; *P* = 0.069).

Selected 500 patients including 100 gastric cancer and 400 non-gastric cancer patients

Matching (n = 500)

AI group (n = 249)

Expert endoscopists group (n = 251)

Included in analysis of per patient diagnosis rate of gastric cancer, invasive and early gastric cancer; per image diagnosis rate of gatric cancer; IOU of gastric cancer among AI diagnosis (n = 249) and expert endoscopists diagnosis (n = 251) groups

▶ **Fig. 1** Study flow diagram.

Outcome	AI diagnosis, 49 patients with gastric cancer with 748 images	Expert endoscopist diagnosis, 51 patients with gastric cancer with 786 images	Risk difference [95 % confidence interval]	
Main outcome				
▪ Per-patient rate of gastric cancer diagnosis	49/49 (100)[1]	48/51 (94.12)	5.88 [−0.58 to 12.3]	
Other outcomes				*P* value
▪ Per-patient rate of invasive gastric cancer diagnosis	22/22 (100)	25/25 (100)	Not applicable	Not applicable
▪ Per-patient rate of early gastric cancer diagnosis	27/27 (100)	23/26 (88.46)	11.54 [−0.74 to 23.82]	0.069
▪ Per-image rate of gastric cancer diagnosis	747/748 (99.87)	693/786 (88.17)	11.7 [9.43 to 13.97]	<0.001
▪ IOU of gastric cancer*, mean ± SD	0.842 ± 0.246	0.972 ± 0.079	−0.13 [−0.15 to −0.11]	<0.001

Abbreviations: AI, artificial intelligence; CNN, convolutional neural network; IOU, intersection over union; SD, standard deviation.
* IOU was evaluated as the area of overlap between the predicted bounding box and the gold-standard bounding box.
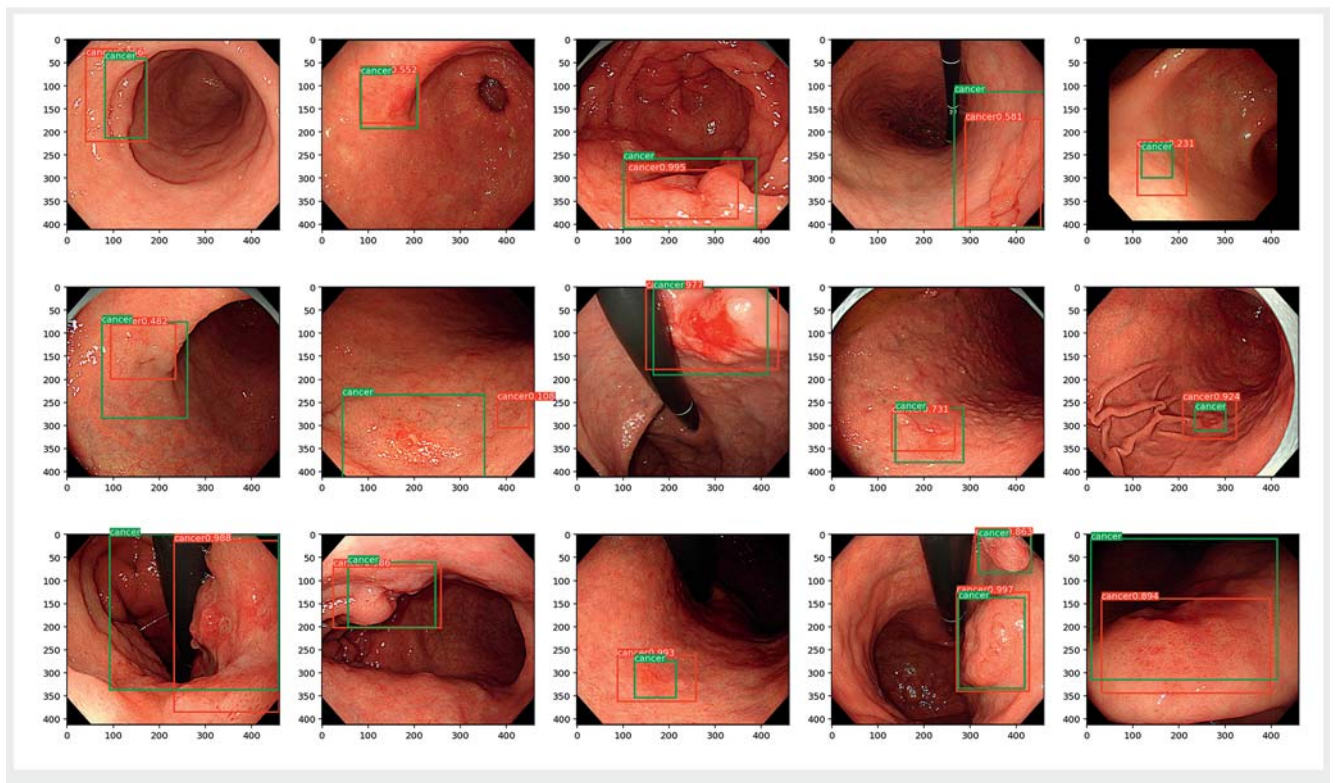
▶ **Fig. 2** Images of gastric cancer used for diagnostic purposes by the artificial intelligence (AI) diagnosis group. Green boxes, gold-standard bounding boxes; red boxes, AI-detected bounding boxes. Source: Keita Otani.

The per-image rate of gastric cancer diagnosis was significantly higher in the AI diagnosis group (747 of 748 images, 99.87 %) than in the expert endoscopist group (693 of 786 images, 88.17 %) (difference 11.7, 95 %CI 9.43 to 13.97; *P* < 0.001).

The IOU of gastric cancer was significantly lower (0.842) in the AI diagnosis group than in the expert endoscopist diagnosis group (0.972) (difference −0.13, 95 %CI −0.15 to −0.11; *P* < 0.001) (▶ **Table 2**, **Table 2 s**).

Discussion

The rate of gastric cancer detection by AI was not inferior to the rate of detection by expert endoscopists. To our knowledge, this study is the first to evaluate patient-level detection rates of early and invasive gastric cancer and to compare AI and expert endoscopists.

The detection rate of AI for gastric cancer was higher than the detection rate of expert endoscopists. We suggest two reasons for this result. First, the per-image rate of gastric cancer diagnosis in the AI diagnosis group was 13.1 % higher than the per-image rate of gastric cancer diagnosis in the expert endoscopist group. A previous study reported a per-image detection rate of gastric cancer of over 96 % [5]; our per-image rate of gastric cancer diagnosis was 99.87 % (747 of 748 images). As the number of images analyzed increased, the likelihood of identifying a cancer increased; this may explain the high detection rate of gastric cancer by AI. Alternatively, the high rate of gastric cancer detection in the AI diagnosis group may be due to the definition of the main outcome, per-patient diagnosis of gastric cancer, as "detected on at least one endoscopic image of gastric cancer." This definition may favor AI diagnosis because AI could suggest many images that potentially include gastric cancer lesions. However, we consider our main outcome to be reasonable when using AI for gastric cancer screening examinations.

The IOU of gastric cancer was significantly lower in the AI diagnosis group (0.09) than in the expert endoscopist group, although the bounding boxes of gastric cancer detected in the AI diagnosis group did not affect the diagnosis of gastric cancer (▶ Fig.2). However, further studies are needed to improve the IOU of gastric cancer by our CNN-based AI diagnosis model.

Our AI model showed a performance in the detection of gastric cancer similar to that of expert endoscopists, even in patients in whom *H. pylori* had been eradicated, who were difficult to evaluate on the basis of endoscopic images [12]. Furthermore, the model was suitable for evaluation of both early and invasive gastric cancers. The AI diagnosis model was developed using 13 584 images of 2639 gastric cancer lesions taken during eight types of endoscopies over a 12-year period [1]. Therefore, our CNN-based AI diagnosis model has potential for use in various patient populations.

This study was the first direct comparison between AI and expert endoscopists of per-patient diagnosis of gastric cancer. However, the study had limitations. First, the study was a single-center retrospective work and potentially affected by selection and confounding bias. Future prospective randomized controlled studies are required. Second, the environment in which images were diagnosed differed from that in which upper endoscopy was performed in practice; this may have compromised the diagnostic accuracy of the expert endoscopists.

In conclusion, we demonstrated noninferiority but not superiority of AI for gastric cancer diagnosis compared with expert endoscopists.

Clinical trial

Trial Registration: ClinicalTrials.gov | Registration number (trial ID): NCT04040374 | Type of study: Retrospective

References

[1] 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

[2] Ogawa R, Nishikawa J, Hideura E et al. Objective assessment of the utility of chromoendoscopy with a support vector machine. J Gastrointest Cancer 2019; 50: 386–391

[3] Ali H, Yasmin M, Sharif M et al. Computer-assisted gastric abnormalities detection using hybrid texture descriptors for chromoendoscopy images. Comput Methods Programs Biomed 2018; 157: 39–47

[4] Sakai Y, Takemoto S, Hori K et al. Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network. Conf Proc IEEE Eng Med Biol Soc 2018; 2018: 4138–4141

[5] Kanesaka T, Lee TC, Uedo N et al. Computer-aided diagnosis for identifying and delineating early gastric cancers in magnifying narrow-band imaging. Gastrointest Endosc 2018; 87: 1339–1344

[6] 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

[7] Lee JH, Kim YJ, Kim YW et al. Spotting malignancies from gastric endoscopic images using deep learning. Surg Endosc 2019; 33: 3790–3797

[8] Riaz F, Silva FB, Ribeiro MD et al. Invariant Gabor texture descriptors for classification of gastroenterology images. IEEE Trans Biomed Eng 2012; 59: 2893–2904

[9] Liu DY, Gan T, Rao NN et al. Identification of lesion images from gastrointestinal endoscope based on feature extraction of combinational methods with and without learning process. Med Image Anal 2016; 32: 281–294

[10] Kubota K, Kuroda J, Yoshida M et al. Medical image analysis: computer-aided diagnosis of gastric cancer invasion on endoscopic images. Surg Endosc 2012; 26: 1485–1489

[11] Zhu Y, Wang QC, Xu MD et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. Gastrointest Endosc 2019; 89: 806–815. e1

[12] Watanabe K, Nagata N, Shimbo T et al. Accuracy of endoscopic diagnosis of Helicobacter pylori infection according to level of endoscopic experience and the effect of training. BMC Gastroenterol 2013; 13: 128