Endoscopy Thieme

Artificial intelligence diagnostic system predicts multiple Lugol-voiding lesions in the esophagus and patients at high risk for esophageal squamous cell carcinoma

Journal:	*Endoscopy*
Manuscript ID	ENDOS-2020-19439.R2
Manuscript Type:	Original Article
Date Submitted by the Author:	22-Oct-2020
Complete List of Authors:	Ikenoyama, Yohei; JFCR, Gastroenterology Yoshio, Toshiyuki; JFCR, Gastroenterology Tokura, Junki; JFCR, Gastroenterology Naito, Sakiko; JFCR, Gastroenterology Namikawa, Ken; JFCR, Gastroenterology Tokai, Yoshitaka; JFCR, Gastroenterology Yoshimizu, Shoichi; JFCR, Gastroenterology Horiuchi, Yusuke; JFCR, Gastroenterology Ishiyama, Akiyoshi; JFCR, Gastroenterology Hirasawa, Toshiaki; JFCR, Gastroenterology Tsuchida, Tomohiro; JFCR, Gastroenterology Katayama, Naoyuki; Mie University Graduate School of Medicine, Hematology and Oncology Tada, Tomohiro; The University of Tokyo, Department of Surgical Oncology Fujisaki, Junko; JFCR, Gastroenterology
Keyword:	Diagnosis and imaging (inc chromoendoscopy, NBI, iSCAN, FICE, CLE) < 01 Endoscopy Upper GI Tract, Image and data processing, documentatiton < 09 Quality and logistical aspects, Training < 09 Quality and logistical aspects
Abstract:	Background and Aims: It is known that an esophagus presenting with multiple Lugol-voiding lesions (LVLs) after iodine staining is high-risk for esophageal cancer; however, it is preferable to identify high-risk cases without iodine staining because iodine causes discomfort and prolongs examination times. In this work, we assessed the capability of an artificial intelligence (AI) system to predict multiple LVLs from non-iodine-stained images as well as high-risk cases for esophageal cancer. Methods: We constructed the AI system by preparing 6634 non-iodine-stained images from 595 cases that underwent endoscopic examination with iodine staining as the training dataset. Its diagnostic capability was then compared with the abilities of 10 experienced endoscopists on an independent validation dataset (667 images from 72 cases). Results: The sensitivity, specificity, and accuracy of each case for the AI system to predict multiple LVLs were 84.4%, 70.0%, and 76.4%, respectively, versus 46.9%, 77.5%, and 63.9%, respectively, for the endoscopists. The AI system had significantly higher sensitivity than 9 out of 10 experienced endoscopists. We found six endoscopic findings that were significantly more frequent in patients with multiple LVLs,

	however the AI system was superior to them to predict multiple LVLs in sensitivity. Moreover, the cases for which the AI system predicted multiple LVLs had significantly more cancers in the esophagus and head and neck than those for which it did not. Conclusion: The AI system could predict multiple LVLs with high sensitivity from non-iodine-stained images. Thus, it will enable endoscopists to apply iodine staining more judiciously.

Production notes

Figs. 1 and 2

Please add lower case figure tags to the individual images, as shown in the image guide; then, please delete the image guide.

Yohei Ikenoyama[1,2], Toshiyuki Yoshio[1,3], Junki Tokura[1], Sakiko Naito[1], Ken Namikawa[1], Yoshitaka Tokai[1], Shoichi Yoshimizu[1], Yusuke Horiuchi[1], Akiyoshi Ishiyama[1], Toshiaki Hirasawa[1,3], Tomohiro Tsuchida[1], Naoyuki Katayama[2], Tomohiro Tada[3,4,5], Junko Fujisaki[1]

[1] Department of Gastroenterology, Cancer Institute Hospital, Japanese Foundation for Cancer Research, Tokyo, Japan

[2] Department of Hematology and Oncology, Mie University Graduate School of Medicine, Mie, Japan

[3] Tada Tomohiro Institute of Gastroenterology and Proctology, Saitama, Japan

[4] AI Medical Service Inc., Tokyo, Japan

[5] Department of Surgical Oncology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Corresponding author

Toshiyuki Yoshio, MD, PhD

Department of Gastroenterology

Cancer Institute Hospital

3-8-31, Ariake

Koto-ku, Tokyo 135-8550

Japan

Email: toshiyuki.yoshio@jfcr.or.jp

Background It is known that an esophagus with multiple Lugol-voiding lesions (LVLs) after iodine staining is high risk for esophageal cancer; however, it is preferable to identify high-risk cases without staining because iodine causes discomfort and prolongs examination times. This study assessed the capability of an artificial intelligence (AI) system to predict multiple LVLs from images that had not been stained with iodine as well as patients at high risk for esophageal cancer.

Methods We constructed the AI system by preparing a training set of 6634 images from white-light and narrow-band imaging in 595 patients before they underwent endoscopic examination with iodine staining. Diagnostic performance was evaluated on an independent validation dataset (667 images from 72 patients) and compared with that of 10 experienced endoscopists

Results The sensitivity, specificity, and accuracy of the AI system to predict multiple LVLs were 84.4%, 70.0%, and 76.4%, respectively, compared with 46.9%, 77.5%, and 63.9%, respectively, for the endoscopists. The AI system had significantly higher sensitivity than 9/10 experienced endoscopists. We also identified six endoscopic findings that were significantly more frequent in patients with multiple LVLs; however, the AI system had greater sensitivity than these findings for the prediction of multiple LVLs. Moreover, patients with AI-predicted multiple LVLs had significantly more cancers in the esophagus and head and neck than patients without predicted multiple LVLs.

Conclusion The AI system could predict multiple LVLs with high sensitivity from images without iodine staining. The system could enable endoscopists to apply iodine staining more judiciously.

Introduction

Esophageal cancer is the seventh most common cancer and the sixth leading cause of cancer deaths, with more than 500 000 deaths per year globally [1]. Esophageal cancer is classified histologically into esophageal squamous cell carcinoma (ESCC), which is common in South America and Asia (including Japan), and adenocarcinoma [1]. Advanced ESCC has a poor prognosis; however, if detected in the early stages, ESCC can be treated with minimally invasive treatments such as endoscopic resection, with a good prognosis [2–4]. Therefore, early detection is very important. However, the detection of superficial ESCC with white-light imaging (WLI) alone is quite difficult, even by esophagogastroduodenoscopy (EGD). Narrow-band imaging (NBI) is useful for detecting superficial ESCC [5–9]; however, it has been reported that even with the use of NBI, inexperienced endoscopists have a low detection rate of only 53% [10].

Chromoendoscopy with Lugol's iodine staining is a useful way to detect ESCC with high sensitivity. However, owing to chest discomfort and prolonged procedure time [11–13], iodine is not usually used in screening EGD, except for very limited cases at high risk for ESCC such as patients with a history of ESCC or head and neck squamous cell carcinoma (HNSCC). It would be more useful if we could identify patients at high risk for ESCC by using endoscopic findings without staining, similarly to how we recognize gastric atrophy as a high-risk finding for gastric cancer during EGD.

When using Lugol's iodine staining as chromoendoscopy, a spotty unstained area is observed in noncancerous epithelium, which we call a Lugol-voiding lesion (LVL). It is well known that patients with multiple LVLs after iodine staining will more frequently have both synchronous and metachronous ESCCs and HNSCCs after endoscopic resection of ESCC [14–18]. Multiple LVLs are associated with heavy smoking and drinking, and a low consumption of green-yellow vegetables [14].

Esophagus with multiple LVLs has been documented with TP53-mutated cells in physiologically normal epithelium, including many precancerous foci as well as multifocal cancers [14,19], also known as the field effect [20,21]. Thus, patients with multiple LVLs are good candidates for targeted screening for ESCCs and HNSCCs by EGD, as they have a high risk for these cancers. However, it is difficult to diagnose an esophagus with multiple LVLs by EGD without iodine chromoendoscopy.

Recently, artificial intelligence (AI) has made remarkable progress in image recognition with deep learning in various medical fields [22–24]. We have also reported the effective application of AI systems in endoscopic diagnosis, such as in detection and invasion depth diagnosis of esophageal cancer [25,26], and detection of gastric [27] and pharyngeal [28] cancer using EGD images.

In this study, we developed an AI diagnostic system to predict the presence of multiple LVLs in the esophagus from EGD images that had not been stained with iodine. The aim of the system was to detect multiple LVLs that could not be detected by endoscopists without iodine staining and to identify patients at high risk of ESCCs and HNSCCs.

Methods

Training dataset

A deep learning-based AI system was developed to predict the presence of multiple LVLs without using Lugol's iodine chromoendoscopy. The system was trained on endoscopic images captured in daily clinical practice at the Cancer Institute Hospital, Tokyo, Japan, from April 2015 to October 2018. Informed consent was obtained from all patients included in the study. All endoscopic images were taken by a high-resolution endoscope (GIF-H290Z; Olympus Medical Systems, Co., Ltd., Tokyo,

Japan) and a high-resolution endoscopic video system (EVIS LUCERA ELITE CV-290/ CLV-290SL; Olympus Medical Systems). The structure enhancement was set to A-mode level 5 for WLI and B-mode level 8 for NBI. Each image was saved as a jpeg file.

For the study, two experienced endoscopists (T.Y. and Y.I.) included non-magnified images of WLI and NBI taken from patients who underwent Lugol's iodine staining (0.75%). We excluded patients with a history of esophagectomy and chemotherapy or radiation to the esophagus. We also excluded images showing esophageal cancer or those of poor quality resulting from poor insufflation, post-biopsy bleeding, halation, blurring, defocus, or mucus. After selection, the two experienced endoscopists classified each image as non-multiple (Grade A/B) or multiple (Grade C) LVLs based on the subsequent Lugol chromoendoscopic images and according to the criteria of Katada et al. (**Fig. 1a–c**) [14], with grade C as an independent indicator of high risk for cancer (Grade A: no LVLs per endoscopic view; Grade B: 1–9 LVLs per endoscopic view; and Grade C: 10 or more LVLs per endoscopic view). Disagreements in diagnosis were resolved throughout discussion until a consensus was reached. These diagnoses were used as the gold standard.

We used these 6634 images from 595 patients as the training set: 3898 images (WLI 1954 images, NBI 1944 images) from 407 patients with non-multiple LVLs (grade A or B), and 2736 images (WLI 1294 images, NBI 1442 images) from 188 patients with multiple LVLs (Grade C). This selection was used as independent images without linking multiple images from the same patient. The training dataset included not only the internal training dataset but also the internal validation dataset. We trained the neural network of our AI system using the internal training dataset and tuned the hyperparameters of the neural network via the internal validation dataset. The

hyperparameters included weight decay, base learning rate, momentum, gamma, and number of iterations. The weight decay reduced the overfitting of the neural network. The learning rate of the neural network was initialized to the base learning rate at the start of training. The momentum was a hyperparameter of the optimizer, and the gamma was the multiplicative factor of the learning rate decay. The parameters of the neural network were updated multiple times, as specified by "number of iterations." We used the settings of weight decay 0.0002, momentum 0.9, base learning rate 0.0001, gamma 0.5, and number of iterations 709900.

AI diagnostic system

We constructed the diagnostic system based on the deep neural network GoogLeNet [29]. GoogLeNet is a convolutional neural network (CNN) consisting of 22 layers. It was the ideal system for developing our dataset because it can classify 1000 classes and can be easily used by most computers. Moreover, using a larger CNN would have made it difficult to suppress overfitting in the CNN learning system from our dataset. The Caffe deep learning framework, originally developed at the Berkeley Vision and Learning Center [30], was then used to train and validate the CNN system.

To optimize our images for GoogLeNet, we resized them to 224 × 224 pixels, and subsequently rotated them for augmentation as preprocessing. We used a pretrained model that learned natural-image features through ImageNet [31]. This procedure, known as transfer learning, is useful even with a small training dataset. In the validation phase, the trained neural network generated a continuous number between zero and one for non-multiple LVLs or multiple LVLs, corresponding to the probability of the condition being present in the image.

Validation of the AI system and endoscopists' diagnosis

To evaluate the diagnostic accuracy of the AI system, an independent validation dataset was prepared. Endoscopic images captured from patients at the Cancer Institute Hospital from November 2018 to July 2019 were collected. Nonmagnified images of WLI and NBI from consecutive patients who also underwent subsequent Lugol's iodine staining were included based on the same criteria as the training set. However, to avoid bias, we did not exclude images of poor quality resulting from poor insufflation, post-biopsy bleeding, halation, blurring, defocus, or mucus. After confirming the selected validation images, the multiple LVLs were also classified by two experienced endoscopists (T.Y. and Y.I.) using the iodine-stained images, and these classifications were used as the gold standard for validation of AI and endoscopists.

The validation dataset included 667 images from 72 patients (WLI 300 images, NBI 367 images), including 325 images (WLI 165 images, NBI 160 images) from 40 patients with non-multiple LVLs and 342 images (WLI 135 images, NBI 207 images) from 32 patients with multiple LVLs (**Fig. 1d–i**). Multiple images from the same patient were presented as a series of image sets for prediction of multiple LVLs. The diagnostic performance of the AI system for predicting multiple LVLs was then evaluated using the validation dataset (see **Fig. 1s** in the online-only Supplementary material).

To compare the diagnostic performance of the AI system with that of endoscopists, 10 board-certified endoscopists from Japan Gastroenterological Endoscopy Society were invited to review the same validation dataset for presence of multiple LVLs. Endoscopists had 8–17 years of experience as doctors and had performed 3500 to 18 000 endoscopic examinations.

Characteristic endoscopic findings to predict multiple LVLs

We selected endoscopic features that would help to identify multiple LVLs, based on discussions of common findings that we had observed during our daily clinical practice and subsequently confirmed on dozens of endoscopic still images. The features of the esophageal mucosa on WLI or NBI that were identified as being characteristic or potentially predictive of multiple LVLs were as follows: 1) few glycogenic acanthosis (<2 per endoscopic image), 2) keratosis, 3) coarse mucosa, 4) invisible mucosal vessels on WLI, 5) reddish background mucosa on WLI, and 6) brownish background mucosa on NBI (**Fig. 2**). Three experienced endoscopists reviewed all the validation set images, evaluated these endoscopic findings in each image, and determined the endoscopists' diagnosis as a majority decision.

Outcome measures

The trained AI system and the 10 board-certified endoscopists determined whether the images of the validation dataset showed non-multiple or multiple LVLs. Endoscopists made a decision for each image, and the majority decision was taken for each patient. The main outcome measures were sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) to predict multiple LVLs. These values were calculated as follows:

$$\text{Sensitivity} = \frac{\text{No. of patients correctly classified with multiple LVLs by AI or endoscopists}}{\text{Total no. of patients with multiple LVLs}}$$

$$\text{Specificity} = \frac{\text{No. of patients correctly classified with non-multiple LVLs by AI or endoscopists}}{\text{Total no. of patients with non-multiple LVLs}}$$

$$\text{PPV} = \frac{\text{No. of patients correctly classified with multiple LVLs by AI or endoscopists}}{\text{No. of patients actually diagnosed with multiple LVLs by AI or endoscopists}}$$

$$\text{NPV} = \frac{\text{No. of patients correctly classified with non-multiple LVLs by AI or endoscopists}}{\text{No. of patients actually diagnosed with non-multiple LVLs by AI or endoscopists}}$$

We retrospectively recorded the number of new ESCCs and HNSCCs detected during regular EGD in patients included in the validation dataset. We included only the cancers detected during the observation period and did not include the trigger cancer for annual (sometimes every 6 months) EGD. We subsequently calculated the incidence of cancers per 100 person–years.

Statistical analysis

Pearson's chi-squared test or Fisher's exact test was used to compare categorical variables of patients and endoscopic findings. A two-sided McNemar test was used to compare the diagnostic performance to predict multiple LVLs between the AI system and the majority decision of the 10 endoscopists. The person–year method was used to calculate the total number of ESCCs and HNSCCs, and to compare the incidence rates per 100 person–years; this measurement considers both the number of patients and the observation period for each patient. The Wald test was used to compare the person–year method. The interobserver agreement among the endoscopists was calculated based on Fleiss' kappa. A P value of <0.05 was considered to indicate statistical significance. All calculations were performed using EZR version 1.27 (Saitama Medical Center, Jichi Medical University, Japan) [32].

Ethics

The study was approved by the Institutional Review Board of the Cancer Institute Hospital (No. 2016–1171) and the Japan Medical Association (ID JMA-IIA00283).

Results

Characteristics of patients in the validation dataset

The characteristics of the patients in the validation dataset are shown in **Table 1**. The ratios of heavy drinkers and current smokers were significantly higher in patients with multiple LVLs than in patients with non-multiple LVLs, whereas there was no difference in age, sex, or flushing reaction between the two groups.

During the observation period, patients with non-multiple LVLs had 5.6 ESCCs and 0.3 HNSCCs per 100 person–years as newly detected cancers, whereas patients with multiple LVLs had 13.3 ESCCs and 4.8 HNSCCs per 100 person–years.

Diagnostic performance of AI and endoscopists

The mean diagnostic times for analyzing the validation dataset of 667 images by the AI system and endoscopists were 60.0 seconds (standard deviation [SD] 0.7) and 121.0 minutes (SD 26.2), respectively. The AI system correctly diagnosed 84.4% (27/32) of patients with multiple LVLs and 70.0% (28/40) of patients with non-multiple LVLs, whereas the experienced endoscopists correctly diagnosed a median 46.9% (15/32) and 77.5% (31/40), respectively. The accuracy of predicting patients with multiple LVLs was 76.4% for the AI system and 63.9% for the experienced endoscopists. The sensitivity of the AI system was significantly higher than that of the experienced endoscopists, whereas the specificity and accuracy were comparable (**Table 2**). The sensitivity of the AI system was significantly higher than that of 9/10 endoscopists. The interobserver agreement value among the endoscopists was 0.264.

Characteristic endoscopic findings to predict multiple LVLs

The findings of few (<2) glycogenic acanthosis per endoscopic image, keratosis, coarse mucosa, reddish background mucosa on WLI, invisible mucosal vessels on WLI, and

brownish background mucosa on NBI were significantly more frequent in patients with multiple LVLs than in those with non-multiple LVLs (**Table 3**).

For all images, the AI system had a sensitivity of 81.6% (279/342) and could predict significantly more multiple LVLs than the findings of few glycogenic acanthosis (<2 per endoscopic image), keratosis, and coarse mucosa (**Fig. 3a**). On WLI images, the AI system had a sensitivity of 81.5% (110/135) and could predict significantly more multiple LVLs than reddish background mucosa (**Fig. 3b**). On NBI images, the AI system had a sensitivity of 81.6% (169/207) and could predict significantly more multiple LVLs than the finding of brownish background mucosa (**Fig. 3c**). The AI system was more sensitive than all endoscopic findings; among the endoscopic findings, the invisible mucosal vessels on WLI resulted in the highest sensitivity (76.3%) for predicting multiple LVLs.

Risk stratification of ESCC and HNSCC by AI diagnostic system

The patients whom the AI system classified as having multiple LVLs had 11.2 ESCCs and 3.4 HNSCCs, resulting in a total of 14.6 ESCCs and HNSCCs per 100 person–years during the observation period. The patients whom the AI system classified as having non-multiple LVLs had 6.1 ESCCs and 0.9 HNSCCs, resulting in a total of 7.0 ESCCs and HNSCCs per 100 person–years during the observation period. The patients whom the AI system classified as having multiple LVLs had significantly more frequent ESCCs ($P < 0.05$) and total ESCCs and HNSCCs ($P < 0.01$) than those with non-multiple LVLs, although the frequency of HNSCCs was not significant ($P = 0.06$). As assumed, the AI system was able to stratify the risk of newly detected cancers as well as detecting existing multiple LVLs.

Discussion

The AI system developed in the current study could predict the presence of multiple LVLs with high sensitivity and could also predict patients at high risk for ESCC and HNSCC from endoscopic images of the esophagus that had not been stained with iodine. The sensitivity of the AI system to predict multiple LVLs was superior to that of experienced endoscopists. To the best of our knowledge, this is the first report of an AI system developed to predict the presence of multiple LVLs and stratify the risk of ESCC and HNSCC.

Conventionally, heavy drinking, smoking, and flushing reaction have been known as risk factors for ESCC [14,33]. Flushing reaction is the blushing of the face typically as a result of drinking one glass of beer in individuals with heterozygous deficiency of aldehyde dehydrogenase 2, causing severe acetaldehydemia [34]. The endoscopic findings of multiple LVLs after iodine staining reflect all these risk factors and stratify the risks of ESCC and HNSCC [14,33]. Multiple LVLs are fairly useful for determining the surveillance schedule after treatment for ESCC or HNSCC; however, we cannot know whether multiple LVLs are present until we conduct iodine staining because there are no widely used findings of EGD to detect multiple LVLs effectively from images that have not been stained with iodine. This means that we usually only know the risk after we have detected cancer or a suspicious cancer lesion. Using this AI system, we can determine the risk for ESCC at the first EGD in every patient after taking some images of the esophagus, which will generalize the concept of risk stratification of multiple LVLs.

In this study, we assessed six endoscopic findings that might predict multiple LVLs from images that have not been stained with iodine. All these findings were significantly more frequent in patients with multiple LVLs. Although the sensitivities

of "few glycogenic acanthosis" and "invisible mucosal vessels" were relatively high, at 72.2% and 76.3%, the sensitivity of the endoscopists was as low as 46.9%, which was not sufficient. This is because the above two features are both negative findings for multiple LVLs (i.e. few glycogenic acanthosis and invisible mucosal vessels). The remaining features identified were not individually useful, as they had low sensitivity. Conversely, the AI system showed a higher sensitivity than each endoscopic finding and compared with experienced endoscopists. The sensitivity of the AI system did not differ between WLI and NBI. Even in the analysis of characteristic endoscopic findings, there was no remarkable finding exclusively from WLI or NBI; unlike cancer detection, there was no difference between WLI and NBI in predicting multiple LVLs.

Matsuno et al. [35] reported that the finding of multiple foci of dilated vessels (MDVs) was useful for predicting multiple LVLs. However, it was difficult to recognize MDVs in non-magnified still images because we had limited information about MDVs, which appeared small and faint. It appears that we would require further training to recognize and use MDVs as a characteristic endoscopic finding of multiple LVLs. In the original report, MDV had high specificity and accuracy; however, the sensitivity was not high, at 55%. To recognize the high-risk cases more easily and avoid missing cancers, we believe that the most important diagnostic parameter is sensitivity, for which the AI system showed the highest value.

The specificity and PPV of the AI system, as well as those of the endoscopists, were low, thus including more false positives. However, in clinical settings, false positives are generally more acceptable than false negatives, as the detection of high-risk individuals would be maximized. Further training using still images and endoscopic videos, in combination with laboratory data, would improve this system. In particular, AI training would be more effective if endoscopic videos that also include many still

images were used. For validation, it would be more useful and convenient if we could predict multiple LVLs from endoscopic videos in clinical practice, allowing automatic feedback without the need to capture still images.

This study has some limitations. First, as this was a single-center retrospective study, we cannot deny the possibility of overfitting, although training data and validation data were completely independent. As a next step, we are considering verification at other institutions. Second, we used a single type of endoscope and endoscopic video system only. Further validation using other endoscopes and systems will generalize this result. Third, patients who had undergone chemoradiation were excluded because they do not show typical multiple LVLs after iodine staining; however, we think that this exclusion was acceptable as the number of such cases was small. Fourth, the study may suffer from bias because the study center specialized in cancer treatment and most of the included patients therefore had some history of cancer, including ESCC and HNSCC, which may lead to a higher incidence of ESCC than in the general population. Fifth, as patients included in the validation dataset only had a short observation period, newly detected ESCCs and HNSCCs may have included cancers that were missed at the index EGD.

In conclusion, we developed an AI system that could predict the presence of multiple LVLs with high sensitivity and identify patients at high risk for cancer using images that had not been stained with iodine. Although the system requires further validation in multicenter and clinical studies, using not only still images but also videos, the current system could enable endoscopists to utilize iodine staining more judiciously.

References

1 Ferlay J, Colombet M, Soerjomataram I et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 2019; 144: 1941–1953

2 Yamashina T, Ishihara R, Nagai K et al. Long-term outcome and metastatic risk after endoscopic resection of superficial esophageal squamous cell carcinoma. Am J Gastroenterol 2013; 108: 544–551

3 Katada C, Muto M, Momma K et al. Clinical outcome after endoscopic mucosal resection for esophageal squamous cell carcinoma invading the muscularis mucosae – a multicenter retrospective cohort study. Endoscopy 2007; 39: 779–783

4 Shimizu Y, Tsukagoshi H, Fujita M et al. Long-term outcome after endoscopic mucosal resection in patients with esophageal squamous cell carcinoma invading the muscularis mucosae or deeper. Gastrointest Endosc 2002; 56: 387–390

5 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: 845–854

6 Takenaka R, Kawahara Y, Okada H et al. Narrow-band imaging provides reliable screening for esophageal malignancy in patients with head and neck cancers. Am J Gastroenterol 2009; 104: 2942–2948

7 Muto M, Minashi K, Yano T et al. Early detection of superficial squamous cell carcinoma in the head and neck region and esophagus by narrow band imaging: a multicenter randomized controlled trial. J Clin Oncol 2010; 28: 1566–1572

8 Kaneko K, Oono Y, Yano T et al. Effect of novel bright image enhanced endoscopy using blue laser imaging (BLI). Endosc Int Open 2014; 2: E212–219

9 Morita FH, Bernardo WM, Ide E et al. Narrow band imaging versus Lugol chromoendoscopy to diagnose squamous cell carcinoma of the esophagus: a systematic review and meta-analysis. BMC Cancer 2017; 17: 54

10 Ishihara R, Takeuchi Y, Chatani R et al. Prospective evaluation of narrow-band imaging endoscopy for screening of esophageal squamous mucosal high-grade neoplasia in experienced and less experienced endoscopists. Dis Esoph 2010; 23: 480–486

11 Sreedharan A, Rembacken BJ, Rotimi O. Acute toxic gastric mucosal damage induced by Lugol's iodine spray during chromoendoscopy. Gut 2005; 54: 886–887

12 Park JM, Lee IS, Kang JY et al. Acute esophageal and gastric injury: complication of Lugol's solution. Scand J Gastroenterol 2007; 42: 135–137

13 Thuler FP, de Paulo GA, Ferrari AP. Chemical esophagitis after chromoendoscopy with Lugol's solution for esophageal cancer: case report. Gastrointest Endosc 2004; 59: 925–926

14 Katada C, Yokoyama T, Yano T et al. Alcohol consumption and multiple dysplastic lesions increase risk of squamous cell carcinoma in the esophagus, head, and neck. Gastroenterology 2016; 151: 860–869

15 Yamashina T, Ishihara R, Nagai K et al. Long-term outcome and metastatic risk after endoscopic resection of superficial esophageal squamous cell carcinoma. Am J Gastroenterol 2013; 108: 544–551

16 Urabe Y, Hiyama T, Tanaka S et al. Metachronous multiple esophageal squamous cell carcinomas and Lugol-voiding lesions after endoscopic mucosal resection. Endoscopy 2009; 41: 304–309

17 Muto M, Takahashi M, Ohtsu A et al. Risk of multiple squamous cell carcinomas both in the esophagus and the head and neck region. Carcinogenesis 2005; 26: 1008–1012

18 Matsubara T, Yamada K, Nakagawa A. Risk of second primary malignancy after esophagectomy for squamous cell carcinoma of the thoracic esophagus. J Clin Oncol 2003; 21: 4336–4341

19 Tian D, Feng Z, Hanley NM et al. Multifocal accumulation of p53 protein in esophageal carcinoma: evidence for field cancerization. Int J Cancer 1998; 78: 568–575

20 Chai H, Brown RE. Field effect in cancer – an update. Ann Clin Lab Sci 2009; 39: 331–337

21 Slaughter DP, Southwick HW, Smejkal W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. Cancer 1953; 6: 963–968

22 Bibault JE, Giraud P, Burgun A. Big data and machine learning in radiation oncology: state of the art and future prospects. Cancer Lett 2016; 382: 110–117

23 Gulshan V, Peng L, Coram M et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016; 316: 2402–2410

24 Esteva A, Kuprel B, Novoa RA et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115–118. Erratum in Nature 2017; 546: 686

25 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

26 Tokai Y, Yoshio T, Aoyama K et al. Application of artificial intelligence using convolutional neural networks in determining the invasion depth of esophageal squamous cell carcinoma. Esophagus 2020; 17: 250–256

27 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

28 Tamashiro A, Yoshio T, Ishiyama A et al. Artificial-intelligence-based detection of pharyngeal cancer using convolutional neural networks. Dig Endosc 2020; 32: 1057–1065

29 Szegedy C, Liu W, Jia Y et al. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015: 1–9. Accessed: March 1 2019. https://arxiv.org/pdf/1409.4842.pdf

30 Jia Y, Shelhamer E, Donahue J et al. Caffe: convolutional architecture for fast feature embedding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014: 1–4. Accessed: March 1 2019; https://arxiv.org/pdf/1408.5093.pdf

31 Deng J, Dong W, Socher R et al. ImageNet: a large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2009: 248–255

32 Kanada Y. Investigation of the freely available easy-to-use software "EZR" for medical statistics. Bone Marrow Transplant 2013; 48: 452–458

33 Yokoyama A, Katada C, Yokoyama T et al. Alcohol abstinence and risk assessment for second esophageal cancer in Japanese men after mucosectomy for early esophageal cancer. PLoS One 2017; 12: e0175182

34 Yokoyama T, Yokoyama A, Kato H et al. Alcohol flushing, alcohol and aldehyde dehydrogenase genotypes, and risk for esophageal squamous cell carcinoma in Japanese men. Cancer Epidemiol Biomarkers Prev 2003; 12: 1227–1233

35 Matsuno K, Ishihara R, Nakagawa K et al. Endoscopic findings corresponding to multiple Lugol-voiding lesions in the esophageal background mucosa. J Gastroenterol Hepatol 2019; 34: 390–396

Fig. 1 Representative images for each grade of Lugol-voiding lesions (LVLs). The endoscopic images are divided into the following three categories according to the number of LVLs. We defined Grade A/B as non-multiple LVLs and Grade C as multiple LVLs. **a** Grade A: no LVLs per endoscopic view. **b** Grade B: 1–9 LVLs per endoscopic view. **c** Grade C: 10 or more LVLs per endoscopic view. **d** White-light imaging (WLI) of LVLs Grade A. **e** WLI of LVLs Grade B. **f** WLI of LVLs Grade C. **g** Narrow-band imaging (NBI) of LVLs Grade A. **h** NBI of LVLs Grade B. **i** NBI of LVLs Grade C.

Fig. 2 Characteristic endoscopic findings in white-light imaging (WLI) and narrow-band imaging (NBI) to classify the grade of Lugol-voiding lesions (LVLs). **a** Glycogenic acanthosis (arrow) in WLI. **b** Glycogenic acanthosis (arrow) in NBI. **c** Keratosis (arrow) in WLI. **d** Keratosis (arrow) in NBI. **e** Coarse mucosa in WLI. **f** Coarse mucosa in NBI. **g** Visible mucosal vessels in WLI. **h** Reddish background mucosa in WLI. **i** Brownish

background mucosa in NBI. Positive findings of c–f, h, i, and negative findings of a, b, g suggest multiple LVLs.

Fig. 3 Sensitivity of the artificial intelligence (AI) system diagnosis and characteristic endoscopic findings to predict multiple Lugol-voiding lesions (LVLs) for each image. The sensitivity of the AI system was significantly higher than that of most endoscopic findings. **a** Sensitivity in all images. **b** Sensitivity in white-light imaging (WLI). **c** Sensitivity in narrow-band imaging (NBI). *Significant difference between two groups ($P < 0.01$).

Table 1 Patient characteristics of the validation dataset.

Characteristics	Non-multiple LVLs (n = 40)	Multiple LVLs (n = 32)	*P* value
Sex, male/female, n	36/4	32/0	0.12
Age, median (range), years	70.5 (48–82)	70 (51–84)	0.67
Alcohol intake[1], never or rarely/light or moderate/heavy, n	7/27/6	2/15/15	<0.05
Flushing, yes/no, n	31/9	26/6	0.78
Smoking, never/former/current, n	7/28/5	4/16/12	<0.05
Person–years	286	210	
Esophagus			<0.01
SCC, n	16	28	
Per 100 person–years	5.6	13.3	
Head and neck			<0.01
SCC, n	1	10	
Per 100 person–years	0.3	4.8	
Esophagus, head and neck			<0.01
SCC, n	17	38	
Per 100 person–years	5.9	18.1	

LVL, Lugol-voiding lesion; SCC, squamous cell carcinoma.

[1]Never or rare, <1 unit/week; light or moderate, <1–17.9 units/week; heavy, ≥18 units/week (1 unit =22 g ethanol).

Table 2 Diagnostic performance to predict patients with multiple Lugol-voiding lesions.

	Sensitivity, % (95%CI)	Specificity, % (95%CI)	PPV, % (95%CI)	NPV, % (95%CI)	Accuracy, % (95%CI)
AI diagnosis	84.4	70.0	69.2	84.8	76.4
Endoscopists' diagnosis (median)	46.9 (40.1–58.7)	77.5 (75.2–80.3)	62.5 (58.6–67.9)	64.6 (62.1–70.4)	63.9 (61.3–69.0)
P value[1]	<0.05	0.15	–	–	0.68

PPV, positive predictive value; NPV, negative predictive value, CI, confidence interval; AI, artificial intelligence.

[1]Comparison between AI diagnosis and experienced endoscopists' diagnosis by majority (McNemar test).

Table 3 Relationship between characteristic endoscopic findings and the grade of Lugol-voiding lesions.

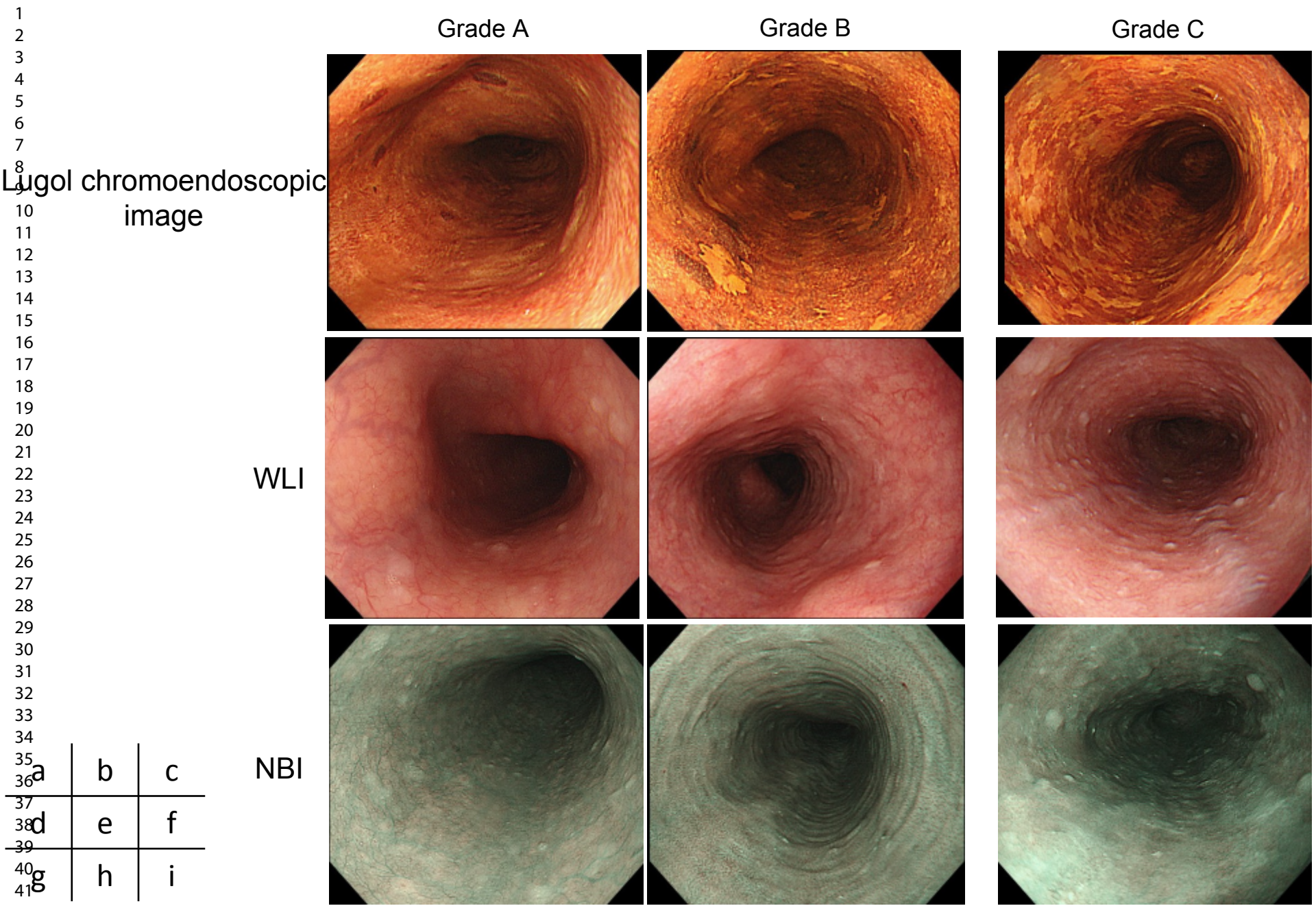
Endoscopic finding	No. of images		*P* value
	Non-multiple LVLs n = 325	Multiple LVLs n = 342	
Few glycogenic acanthosis[1] (+/–)	122/203	247/95	<0.01
Keratosis (+/–)	27/298	125/217	<0.01
Coarse mucosa (+/–)	41/284	177/165	<0.01
Reddish background mucosa in WLI (+/–)	14/151	48/87	<0.01
Invisible mucosal vessels in WLI (+/–)	92/73	103/32	<0.01
Brownish background mucosa in NBI (+/–)	9/151	84/123	<0.01

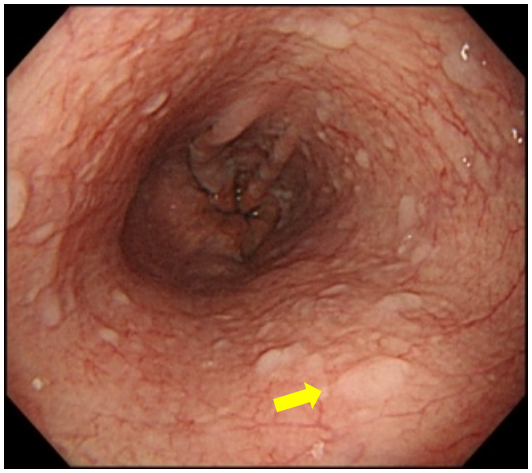
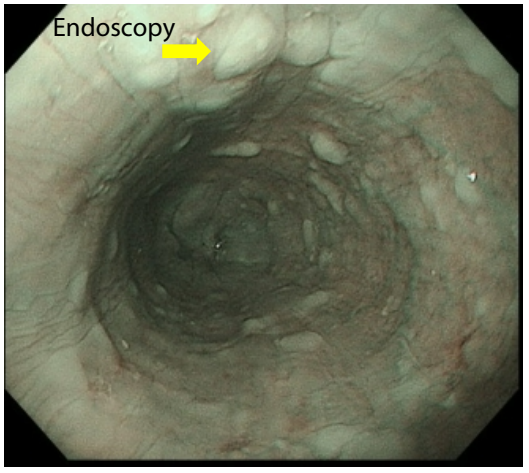
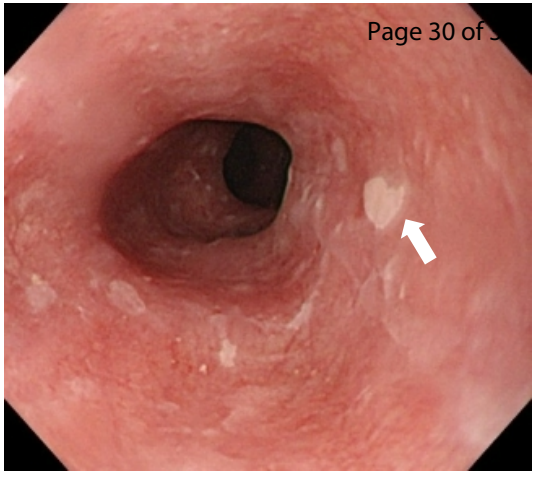
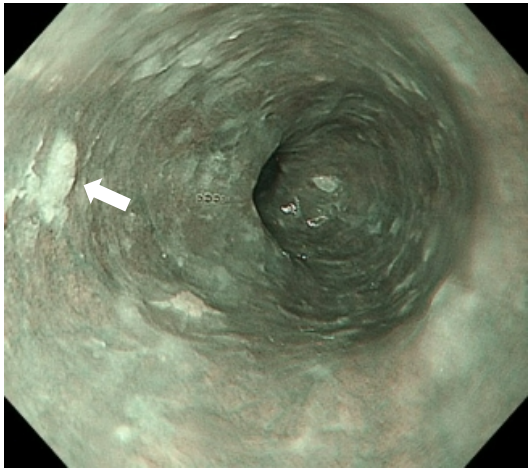
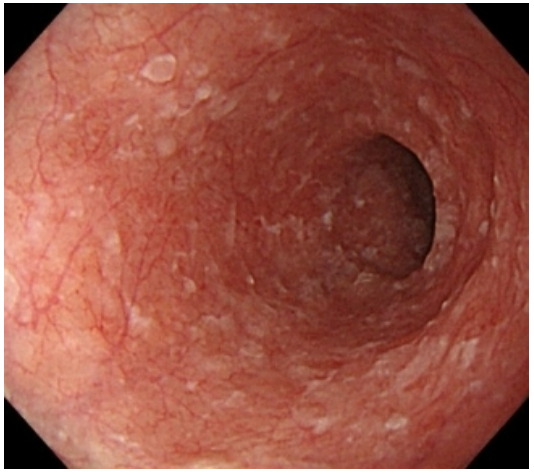
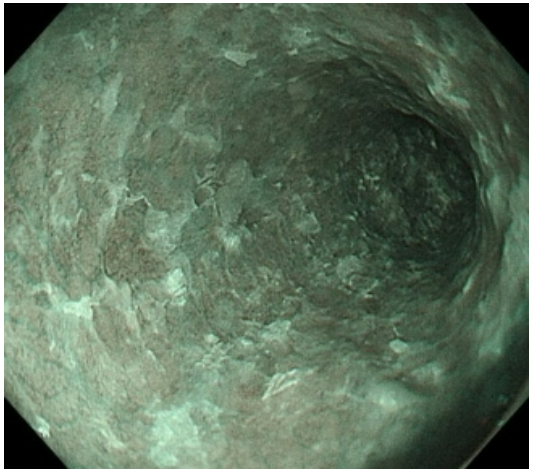
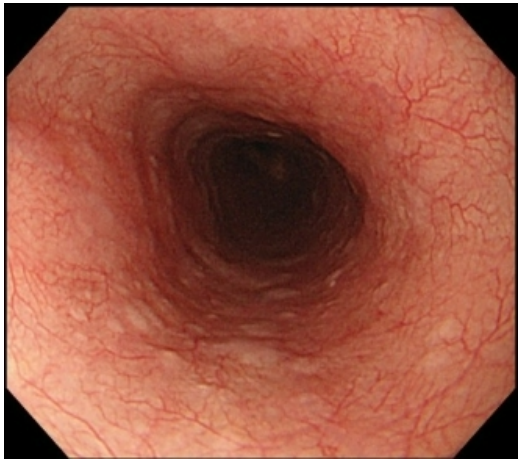
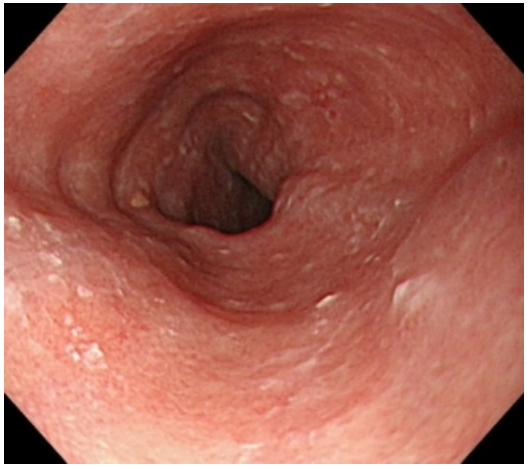
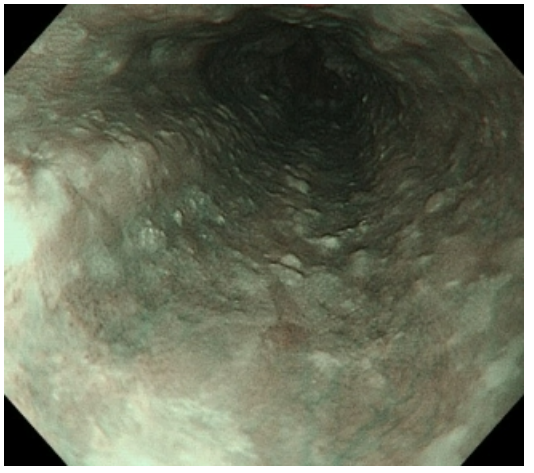
LVL, Lugol-voiding lesion; WLI, white-light imaging; NBI, narrow-band imaging.

[1]<2 per endoscopic view.

In brief

This study developed and evaluated an artificial intelligence (AI) system for predicting the presence of multiple Lugol-voiding lesions (LVLs) from images that had not been stained with iodine. The AI system achieved significantly higher sensitivity rates than a group of 10 experienced endoscopists. This finding is important as it could indirectly lead to higher detection rates of (pre-)cancerous lesions in the esophagus and head and neck region in screening programs.


Non-multiple LVLs
Multiple LVLs
Grade A
Grade B
Grade C
Lugol chromoendoscopic image
WLI
NBI
a
b
c
d
e
f
g
h
i

a	b	c
d	e	f
g	h	i

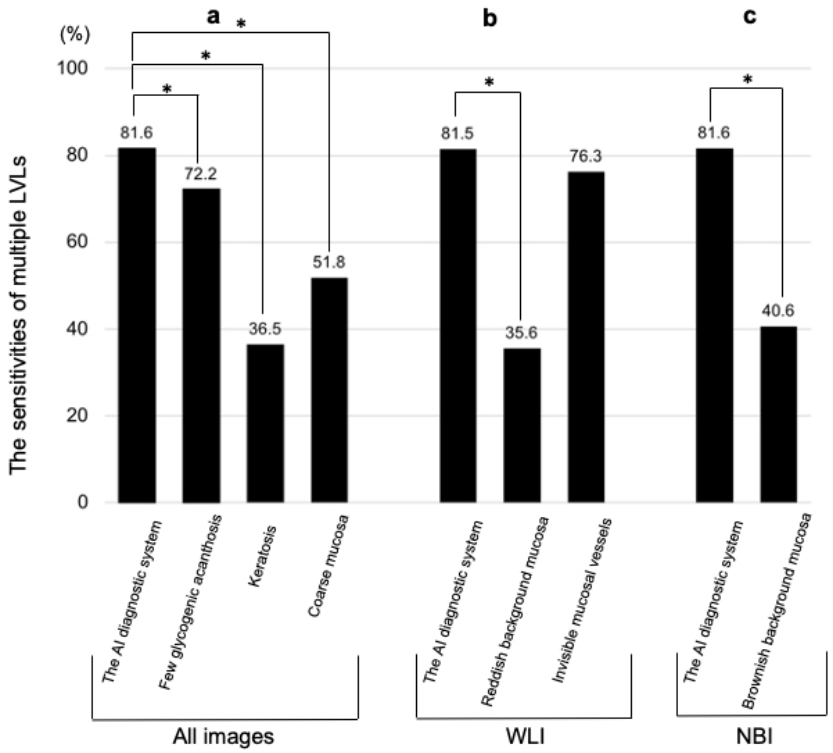


Figure 3. The sensitivity of the AI system diagnosis and characteristic endoscopic findings to predict multiple LVLs for each image.

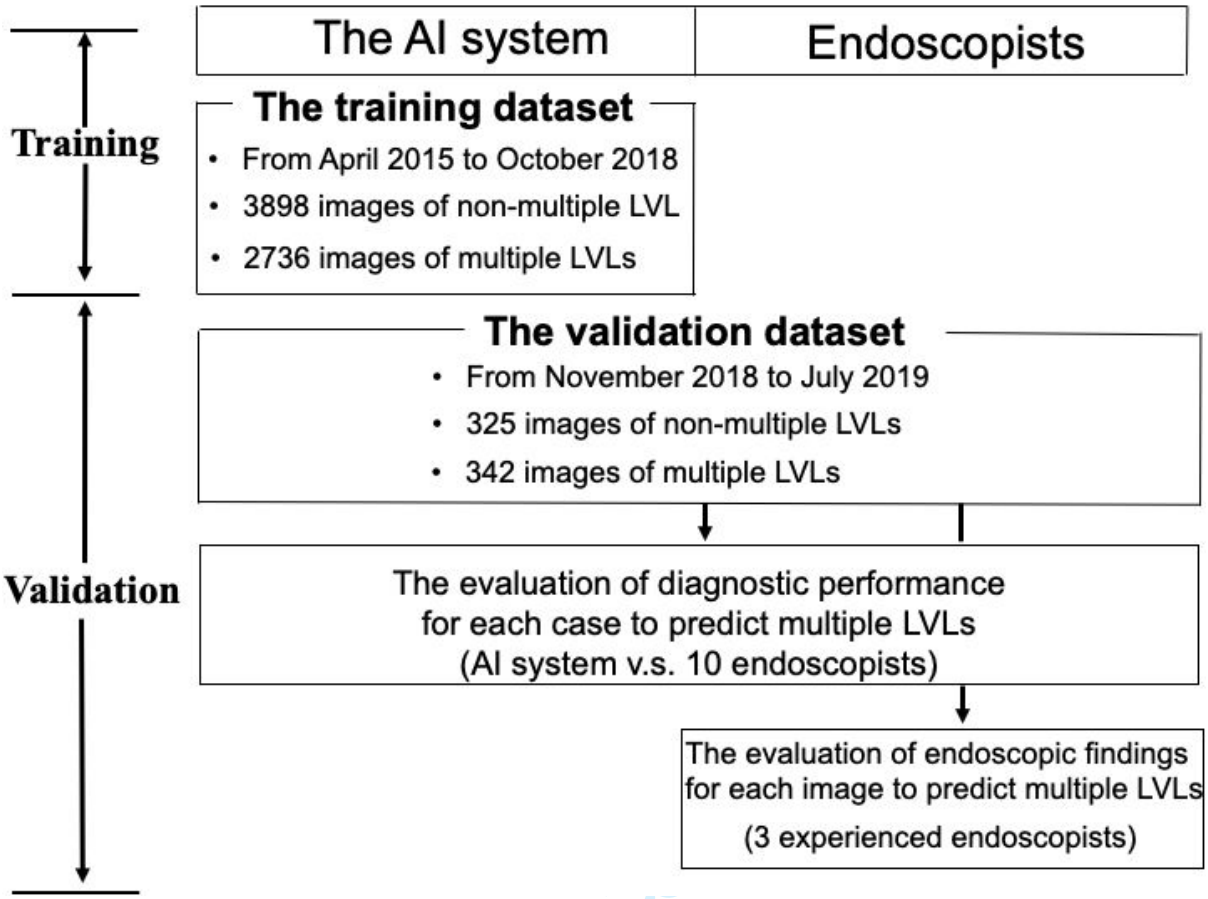
254x190mm (72 x 72 DPI)

Artificial intelligence diagnostic system predicts multiple Lugol-voiding lesions in the esophagus and patients at high risk for esophageal squamous cell carcinoma

Yohei Ikenoyama1

Supplementary material

Fig. 1s Flow chart of the study design.



TRIPOD Checklist: Prediction Model Development

Section/Topic	Item	Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	3
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	5,6
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	5,6
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	7-11
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	7
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	7
	5b	Describe eligibility criteria for participants.	7,8
	5c	Give details of treatments received, if relevant.	NA
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	12
	6b	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	7-10
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	Explain how the study size was arrived at.	NA
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	NA
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	12
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	8,9
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	12
Risk groups	11	Provide details on how risk groups were created, if done.	NA
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	13,14
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	13,14
Model development	14a	Specify the number of participants and outcome events in each analysis.	10,11
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	NA
	15b	Explain how to the use the prediction model.	9,10
Model performance	16	Report performance measures (with CIs) for the prediction model.	14,15
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	18,19
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	16-18
Implications	20	Discuss the potential clinical use of the model and implications for future research.	18,19
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	19,20
Funding	22	Give the source of funding and the role of the funders for the present study.	NA

We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.