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

Authors
Yohei Ikenoyama1,2, Toshiyuki Yoshio1,3, Junki Tokura1, Sakiko Naito1, Ken Namikawa1, Yoshitaka Tokai1, Shoichi Yoshimizu1, Yusuke Horiechi1, Akiyoshi Ishiyama1, Toshiaki Hirasawa1,3, Tomohiro Tsuchida1, Naoyuki Katayama2, Tomohiro Tada3,4,5, Junko Fujisaki1

Institutions
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

Submitted 4.6.2020
Accepted after revision 20.11.2020
Published online 2021

ABSTRACT
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 Kataeda et al. [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 augmenta-
tion 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. Non-magnified 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 de-
cision 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

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

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 1 Patient characteristics of the validation dataset.
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 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 inter-observer 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 on WLI (Fig. 3b). On NBI images, the AI system had a sensitivity of 81.6% (169/207) and could predict significantly more multiple LVLs than 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.

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

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity, % (95 %CI)</th>
<th>Specificity, % (95 %CI)</th>
<th>PPV, % (95 %CI)</th>
<th>NPV, % (95 %CI)</th>
<th>Accuracy, % (95 %CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI diagnosis</td>
<td>84.4</td>
<td>70.0</td>
<td>69.2</td>
<td>84.8</td>
<td>76.4</td>
</tr>
<tr>
<td>Endoscopists’ diagnosis (median)</td>
<td>46.9 (40.1 – 58.7)</td>
<td>77.5 (75.2 – 80.3)</td>
<td>62.5 (58.6 – 67.9)</td>
<td>64.6 (62.1 – 70.4)</td>
<td>63.9 (61.3 – 69.0)</td>
</tr>
<tr>
<td>P value</td>
<td>&lt;0.05</td>
<td>0.15</td>
<td>–</td>
<td>–</td>
<td>0.68</td>
</tr>
</tbody>
</table>

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.

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

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

1 < 2 per endoscopic view.
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 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.

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).
Patients with 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.

**Funding**

E-DA Hospital the Daiwa Securities Health Foundation, 3119 Grant the Takeda Science Foundation Grant, JSPS KAKENHI the Uehara Memorial Foundation Grant 19K08408

**Competing interests**

Tomohiro Tada is a shareholder in AI Medical Service, Inc. All other authors declare that they have no conflicts of interest.

**References**


Erratum in Nature 2017; 546: 686


