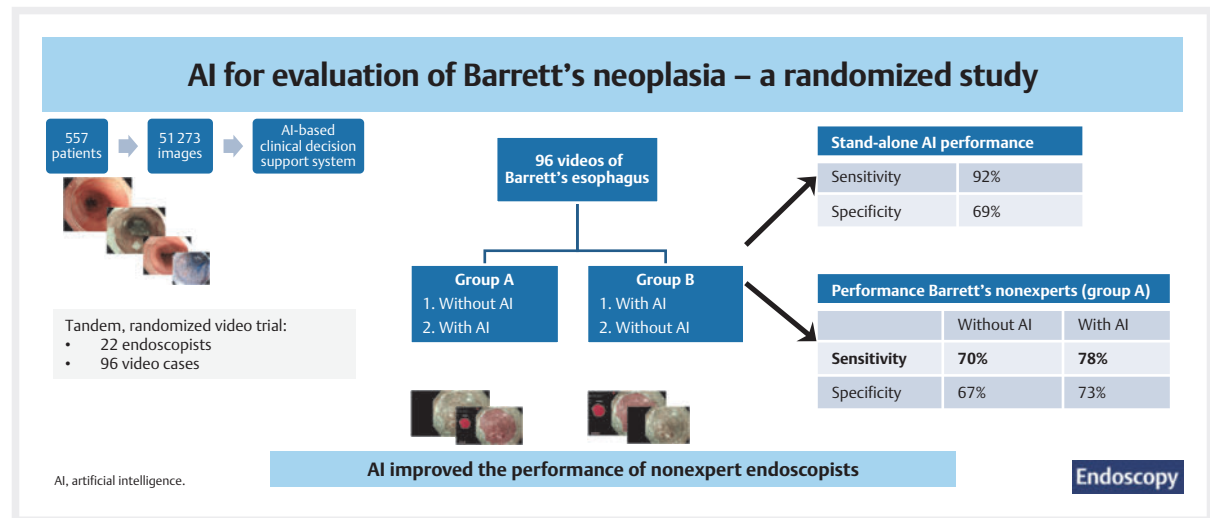


# Influence of artificial intelligence on the diagnostic performance of endoscopists in the assessment of Barrett's esophagus: a tandem randomized and video trial

## GRAPHICAL ABSTRACT



### Authors

Michael Meinikheim<sup>1,†</sup>, Robert Mendel<sup>2,‡</sup>, Christoph Palm<sup>2</sup>, Andreas Probst<sup>1</sup>, Anna Muzalyova<sup>1</sup>, Markus W. Scheppach<sup>1</sup>, Sandra Nagl<sup>1</sup>, Elisabeth Schnoy<sup>1</sup>, Christoph Römmele<sup>1</sup>, Dominik A. H. Schulz<sup>1</sup>, Jakob Schlottmann<sup>1</sup>, Friederike Prinz<sup>1</sup>, David Rauber<sup>2</sup>, Tobias Rückert<sup>2</sup>, Tomoaki Matsumura<sup>3</sup>, Glòria Fernández-Esparrach<sup>4,5,6,7</sup>, Nasim Parsa<sup>8,9</sup>, Michael F. Byrne<sup>9,10</sup>, Helmut Messmann<sup>1,\*</sup>, Alanna Ebigbo<sup>1,\*</sup>

### Institutions

- 1 Department of Gastroenterology, University Hospital Augsburg, Augsburg, Germany
- 2 Regensburg Medical Image Computing, Ostbayerische Technische Hochschule Regensburg, Regensburg, Germany
- 3 Department of Gastroenterology, Chiba University Graduate School of Medicine, Chiba, Japan
- 4 Endoscopy Unit, Gastroenterology Department, ICMDM, Hospital Clínic de Barcelona, Barcelona, Spain
- 5 Faculty of Medicine, University of Barcelona, Barcelona, Spain
- 6 Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain
- 7 Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Barcelona, Spain

- 8 Division of Gastroenterology and Hepatology, Mayo Clinic, Scottsdale, United States
- 9 Satisfai Health, Vancouver, Canada
- 10 Gastroenterology, Vancouver General Hospital, The University of British Columbia, Vancouver, Canada

received 19.9.2023

accepted after revision 13.3.2024

accepted manuscript online 28.3.2024

published online 2024

### Bibliography

Endoscopy

DOI 10.1055/a-2296-5696


ISSN 0013-726X

© 2024. The Author(s).

This is an open access article published by Thieme under the terms of the Creative Commons Attribution-NonDerivative-NonCommercial License, permitting copying and reproduction so long as the original work is given appropriate credit. Contents may not be used for commercial purposes, or adapted, remixed, transformed or built upon. (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)  
Georg Thieme Verlag KG, Rüdigerstraße 14,  
70469 Stuttgart, Germany

† These authors contributed equally.

\* Co-senior authors.

 **Supplementary Material**  
Supplementary Material is available under  
<https://doi.org/10.1055/a-2296-5696>

#### Corresponding author

Michael Meinikheim, MD, Department of Gastroenterology,  
University Hospital Augsburg, Stenglinstraße 2, 86156  
Augsburg, Germany  
[michael.meinikheim@med.uni-augsburg.de](mailto:michael.meinikheim@med.uni-augsburg.de)

#### ABSTRACT

**Background** This study evaluated the effect of an artificial intelligence (AI)-based clinical decision support system on the performance and diagnostic confidence of endoscopists in their assessment of Barrett's esophagus (BE).

**Methods** 96 standardized endoscopy videos were assessed by 22 endoscopists with varying degrees of BE experience from 12 centers. Assessment was randomized into two video sets: group A (review first without AI and second with AI) and group B (review first with AI and second without AI). Endoscopists were required to evaluate each video

for the presence of Barrett's esophagus-related neoplasia (BERN) and then decide on a spot for a targeted biopsy. After the second assessment, they were allowed to change their clinical decision and confidence level.

**Results** AI had a stand-alone sensitivity, specificity, and accuracy of 92.2%, 68.9%, and 81.3%, respectively. Without AI, BE experts had an overall sensitivity, specificity, and accuracy of 83.3%, 58.1%, and 71.5%, respectively. With AI, BE nonexperts showed a significant improvement in sensitivity and specificity when videos were assessed a second time with AI (sensitivity 69.8% [95%CI 65.2%–74.2%] to 78.0% [95%CI 74.0%–82.0%]; specificity 67.3% [95%CI 62.5%–72.2%] to 72.7% [95%CI 68.2%–77.3%]). In addition, the diagnostic confidence of BE nonexperts improved significantly with AI.

**Conclusion** BE nonexperts benefitted significantly from additional AI. BE experts and nonexperts remained significantly below the stand-alone performance of AI, suggesting that there may be other factors influencing endoscopists' decisions to follow or discard AI advice.

## Introduction

Barrett's esophagus (BE) is a precursor of esophageal adenocarcinoma. Although studies suggest that the rate of progression of nondysplastic Barrett's esophagus (NDBE) to Barrett's esophagus-related neoplasia (BERN) is low, once dysplasia is present, the risk of progression increases significantly [1]. Recent data have demonstrated an increase in the incidence of esophageal adenocarcinoma in the Western world [2, 3]. Early detection of esophageal adenocarcinoma determines the patient's prognosis [4]. During endoscopy, BERN is difficult to detect and often challenging to distinguish from NDBE. Miss rates of more than 20% for BERN demonstrate that existing strategies for dysplasia detection may need improvement [5].

Artificial intelligence (AI) has undergone intense research in endoscopy, with numerous potential applications [6]. One possibility for AI is to offer a "second opinion" or decision support during the endoscopic evaluation of BE. Several research teams have used deep learning to develop AI-based clinical decision support systems for computer-aided detection (CADe) and computer-aided diagnosis (CADx) in the context of BE assessment and BERN [7, 8, 9, 10, 11, 12, 13]. Although existing trials have shown promising results regarding sensitivity, specificity, and accuracy, performance measures refer mostly to CADe or CADx on still images [8, 10, 11, 14]. Moreover, most trials have evaluated the stand-alone performance of an AI system and compared it with the stand-alone performance of endoscopists rather than investigating the add-on effect of AI on the performance of endoscopists, as described by the position statement of the European Society of Gastrointestinal Endoscopy (ESGE) [15].

Most screening and surveillance endoscopic examinations of BE are conducted in an outpatient setting and by endoscopists who are non-BE experts. In line with the ESGE statements on the expected value of AI, we sought to investigate the effects an AI system has on the performance of BE nonexpert endoscopists assessing a Barrett's video dataset.

## Methods

A multicenter, randomized, controlled tandem video trial was conducted to evaluate the add-on effect of an AI system on the performance of endoscopists during the evaluation of BE. We implemented the DECIDE-AI guidelines for reporting our study results [16].

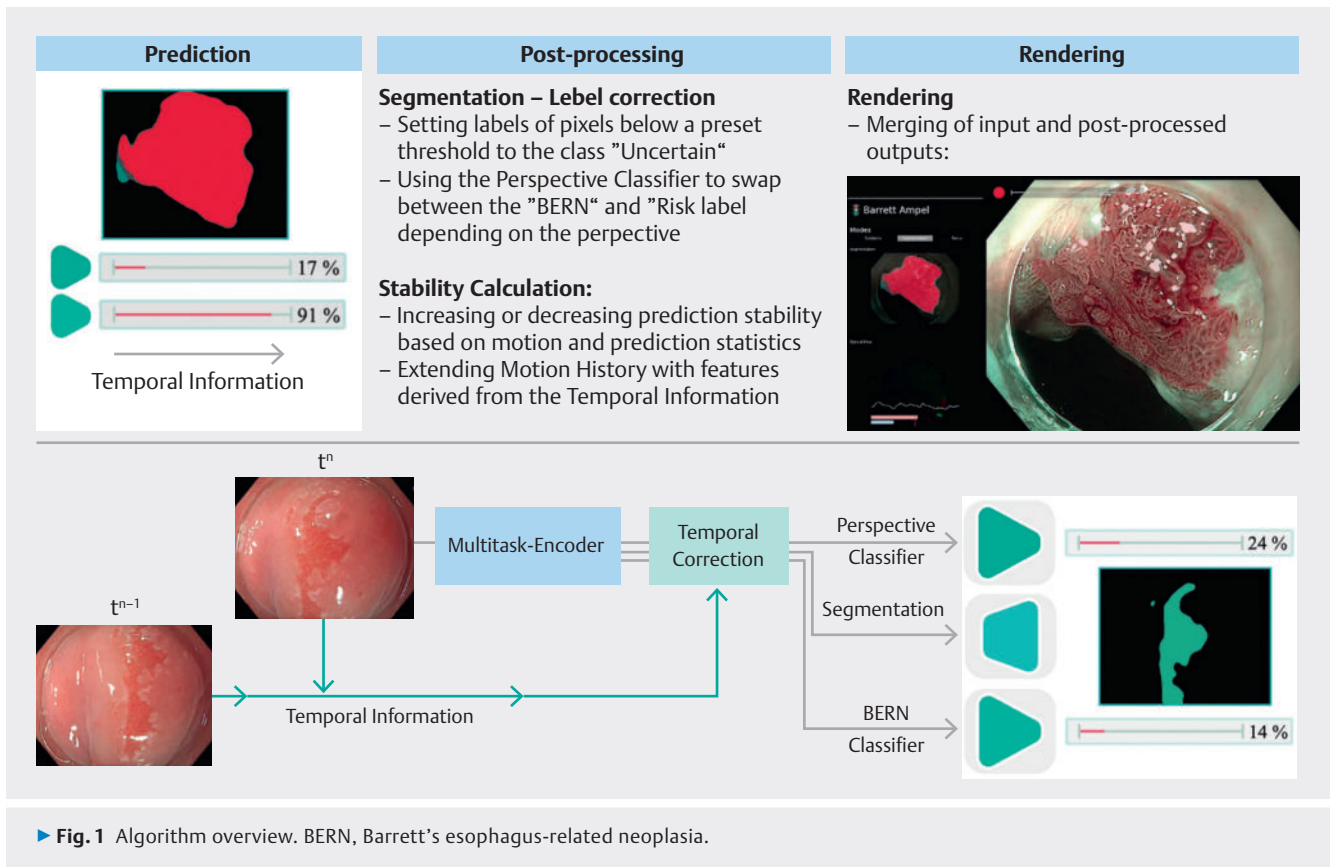
### Study outcomes

The primary outcome was the effect of AI on the diagnostic performance of nonexpert endoscopists in BE evaluation. Secondary outcomes included: 1) the stand-alone performance of AI for the detection and segmentation of BERN; 2) the effect of AI on the diagnostic performance of expert endoscopists in BE evaluation; 3) the effect of AI on the diagnostic confidence of expert and nonexpert endoscopists in BE evaluation.

## Development of the AI system

### Training data

The training dataset included overview and near-focus images of the region of interest (ROI) in high definition white-light endoscopy, narrow-band imaging, texture and color enhancement imaging, as well as chromoendoscopy with acetic acid



and indigo carmine. The complete dataset consisted of images from 557 patients, including 51 273 images.

The fully labeled portion of the dataset included images from 456 patients, 152 with NDBE, and 304 with BERN. This data pool consisted of 3210 labeled training images. All images were assessed by BE expert endoscopists and histologically confirmed. In addition to image-level classification, a pixel-level segmentation was prepared by BE expert endoscopists. For the pixel-level labels, the experts delineated normal tissue, NDBE, BERN, and regions at risk. Areas labeled "at risk" show histologically confirmed BERN from a distance or perspective that does not allow an accurate visual assessment. More detailed descriptions are presented in **Methods 1s** in the online-only Supplementary material.

### Deep learning model

The deep learning model was based on the DeepLabV3+ [17] architecture with kernel-sharing [18] and a ResNet50 [19] backbone.

The segmentation task was trained with the semi-supervised Error-Correcting Mean-Teacher [20] algorithm. More detailed descriptions are presented in **Methods 2s**.

### Algorithm

The algorithm integrates information into the trained model to provide consistent predictions. ► **Fig. 1** shows a comprehensive overview of the components involved. Both the predicted motion of the incoming endoscopic data, as well as the stability of

the model's prediction influence an internal counting algorithm. Only when both parts pass a stability threshold are the model predictions marked on the screen. More detailed descriptions are presented in the **Methods 3s**.

### Description of the video trial data

A total of 96 prospectively collected videos of endoscopic examinations in 72 consecutive patients who presented to the University Hospital of Augsburg for evaluation of BE and BERN between 1 October 2021 and 30 September 2022, and who met the study criteria, were included. Included patients were either referred for further evaluation of BE/BERN or presented for surveillance of BE. Informed consent of all patients was ensured. Approval by the ethics committee of the Ludwig Maximilians University of Munich was granted (PNO: 20–010).

We included overview and close-up videos with a duration of between 15 seconds and 90 seconds. Although most videos showed the entire BE segment, some videos showed only a portion of the esophagus. We included 45 cases of NDBE (46.9%), five cases of low grade dysplasia (LGD; 5.2%), seven cases of high grade dysplasia (HGD; 7.3%), 36 cases of T1a adenocarcinoma (37.5%), and three cases of T1b adenocarcinoma (3.1%). BERN lesions included in this trial was exclusively flat or slightly elevated (Paris IIa/IIb) (► **Table 1**).

All included cases contained at least two imaging modalities, including high definition white-light endoscopy, narrow-band imaging, or texture and color enhancement imaging. Data were obtained from endoscopic examinations with Olym-

► **Table 1** Distribution of histology and length of Barrett's esophagus segment.

	Length of BE segment			Total
	2–3 cm	3–10 cm	≥10 cm	
Histology, n (%)				
▪ NDBE	22 (22.9)	18 (18.8)	5 (5.2)	45 (46.9)
▪ LGD	3 (3.2)	1 (1.0)	1 (1.0)	5 (5.2)
▪ HGD	2 (2.1)	5 (5.2)	0	7 (7.3)
▪ T1a	19 (19.8)	16 (16.7)	1 (1.0)	36 (37.5)
▪ T1b	1 (1.0)	2 (2.1)	0	3 (3.1)

BE, Barrett's esophagus; NDBE, nondysplastic Barrett's esophagus; LGD, low grade dysplasia; HGD, high grade dysplasia; T1a and T1b, according to TNM classification of malignant tumors.

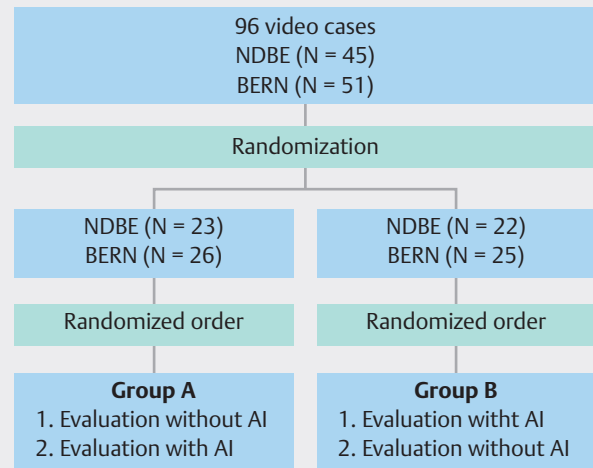
pus GIF-HQ190, GIF-XZ1200, GIF-EZ1500 gastroscopes, and CV-1500 Evis X1 endoscopic processor (Olympus, Tokyo, Japan). Video documentation of forceps biopsy or endoscopic resection of an ROI was performed to enable correlation of histological assessment (ground truth) with the endoscopic assessment. Video cases were included only where histological proof of the ROI was available. If more than one video case from the same patient was included, the videos were taken in a way that no visual overlap occurred between the video cases. Histological assessment was performed by pathologists specialized in BE assessment, and a second, independent pathologist always confirmed the results.

### Design of the trial

To evaluate whether the additional use of AI affects the performance of endoscopists with varying levels of expertise, a tandem study design was chosen. To this end, video cases were demonstrated twice – with and without additional AI. This means that 50% of cases were presented to the study participants first without and second with additional AI (group A). The other half of the cases were presented in the opposite order, first with and second without additional AI support (group B). In addition, the dataset was divided into cases of BERN and NDBE. Within these groups, we conducted a permuted block randomization (1:1) of the allocation to either group A or group B. Finally, the resulting subgroups of NDBE and BERN were again combined, and the order of appearance was randomized to create the final test set (► Fig. 2).

### Evaluation of the influence of AI on diagnostic confidence

For each video, participants indicated their level of confidence on a scale from 0 to 9. Confidence levels were divided into two basic groups: “low confidence” for 0–4 and “high confidence” for 5–9 regarding how sure or unsure participants were of their diagnoses.



► **Fig. 2** Schematic representation of the test set for the participants of the trial. NDBE, nondysplastic Barrett's esophagus; BERN, Barrett's esophagus-related neoplasia; AI, artificial intelligence.

### Statistical analysis

Sensitivity was defined as the correct diagnosis of video cases with neoplasms and, at the same time, the correct localization of neoplasia with a digital biopsy spot within the video case. The ground truth was expert assessment, which was confirmed histologically. Specificity was defined as the correct diagnosis of video cases without a visible neoplasm as NDBE.

Based on previous work [14, 21, 22, 23], the sensitivity of general endoscopists without particular BE experience and without the support of AI was estimated to be approximately 60%. With the support of AI, sensitivity was estimated to be 80%. We invited consecutive patients referred for evaluation or surveillance who met the inclusion criteria during the period from 1 October 2021 to 30 September 2022. As described above, 96 video cases were then generated from these 72 included patients, making sure to avoid video overlaps within the same patient.

Performance metrics of the study participants, including sensitivity, specificity, and accuracy are presented as percentages. As the performance of each group with and without AI was captured on the same set of videos and thus represented paired samples, results were tested for statistically significant differences using McNemar's test. We used Wald interval as the method to determine the confidence intervals. The performance of nonexpert endoscopists with additional AI was compared with the benchmark performance of Barrett's experts and tested for statistically significant differences using the chi-squared test. In addition, differences in performance depending on the confidence level were tested using the chi-squared test. The significance level was set at 0.05. All statistical tests were performed using SPSS version 28.0 (IBM Corp., Armonk, New York, USA).

## Endoscopists

The aim was to recruit BE experts as well as BE nonexpert endoscopists. Overall, 33 endoscopists (12 BE experts and 21 BE nonexperts) were invited to participate in the trial. Finally, 22 participants (six BE experts and 16 nonexperts) from four countries and 12 institutions, including six hospitals and six private practices, completed the video trial. A detailed description of the participating endoscopists is shown in **Table 1s**. BE experts were defined according to the position statement of the ESGE, including endoscopists with regular BE evaluation and with experience of at least 30 BERN resections and 30 endoscopic ablations [24]. Nonexperts were board-certified gastroenterologists who did not meet the criteria of experienced endoscopists in the context of BE. Nonexpert endoscopists were further subdivided into three groups: endoscopists in private practices, endoscopists in secondary care hospitals, and nonexperts working in BE referral centers.

## Trial framework

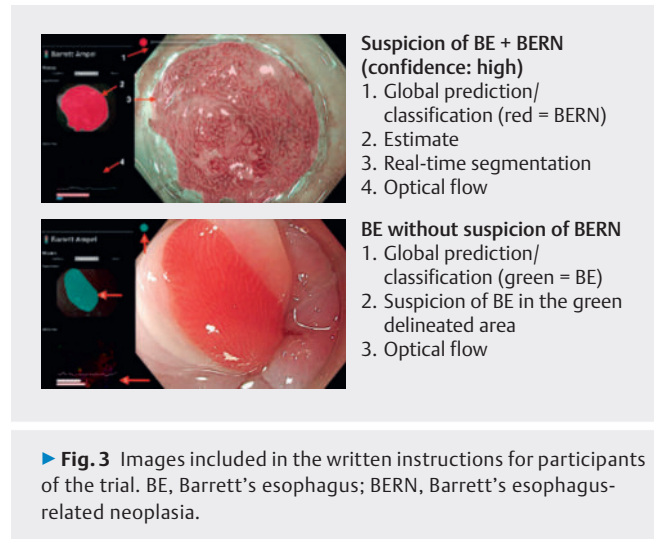
Participants conducted the online video trial with a dedicated software tool specifically designed for this study (**Methods 4s**). The fully anonymized video test set was displayed to participants in a predetermined order. Participants were asked to classify each video for the presence or absence of BERN. When a BERN was assumed, participants were required to include a single spot for a targeted biopsy. No biopsy spot was demarcated for video cases without an assumption of BERN; such videos were left unaltered. Each video could be reassessed as often as the participants wished; however, it was no longer possible to return to the previous video after proceeding to the next video. For every case, participants had to indicate their confidence level in the correctness of their diagnosis before moving to the next video.

The output of AI (global prediction, segmental overlay) was dynamic; this means that the information produced by AI on the video screen was not always continuous and changed with the position of the scope or the region of endoscopic focus. We differentiated between stable and nonstable predictions to evaluate the persistency of a prediction by AI. Stable prediction was defined as a segmentation heat map displayed for more than 3 seconds (150 consecutive frames). Nonstable prediction implied cases where the segmentation map repeatedly appeared at the same spot for an overall cumulative time of more than 3 seconds (150 frames) but not continuously (► **Fig. 3**).

## Results

### Evaluation of the add-on effect of AI on the performance of nonexpert endoscopists

When participants were evaluated initially without AI and subsequently with AI (group A), they improved their sensitivity from 69.8% (95%CI 65.2%–74.2%) to 78.0% (95%CI 74.0%–82.0%) and their specificity from 67.3% (95%CI 62.5%–72.2%) to 72.7% (95%CI 68.2%–77.3%). When the initial evaluation was done with AI (group B), the performance of nonexperts



did not change (sensitivity of 73.1% [95%CI 68.8%–77.4%] and 73.1% [95%CI 68.8%–77.4%]; specificity of 60.3% [95%CI 55.2%–65.2%] and 61.1% [95%CI 56.2%–66.3%]) (► **Table 2, Fig. 1s**).

Participants from secondary care hospitals improved their performance with AI, but this difference did not reach statistical significance (► **Table 2**).

Gastroenterologists from private practices benefitted significantly from additional AI, with a sensitivity improvement from 62.0% (95%CI 54.0%–69.3%) to 74.7% (95%CI 67.3%–81.3%) and an accuracy improvement from 67.7% (95%CI 62.1%–73.0%) to 75.2% (95%CI 70.2%–80.1%) in group A (► **Table 2**). There was no significant improvement in sensitivity, specificity, and accuracy in group B (AI first).

For nonexperts in BE referral centers, sensitivity improved significantly from 78.7% (95%CI 72.0%–85.3%) to 85.3% (95%CI 79.3%–90.7%) in group A. In group B, the performance was not significantly different after the first review with AI (► **Table 2**).

### Stand-alone performance of AI

AI classified 47/51 videos with BERN lesions correctly (sensitivity of 92.2% [95%CI 88.2%–95.6%]), while 31/45 videos without BERN were classified correctly as ND BE (specificity of 68.9% [95%CI 62.2%–75.6%]) (► **Table 3**). The system's overall accuracy on this test set was 81.3% (95%CI 77.3%–85.2%). In 39/47 correctly classified cases, the lesion was precisely detected, and the respective lesion's segmentation overlay appeared for at least 150 frames on the main screen (stable prediction). The global classification correctly predicted the video as BERN in eight cases, but the segmentation overlay persisted for fewer than 150 frames on the respective lesion (nonstable prediction). False-positive results appeared in 14 cases (six nonstable predictions with false-positive segmental overlays of fewer than 150 frames, and eight stable predictions with false-positive segmental overlays of more than 150 frames). In four cases (1 LGD, 1 HGD, 2 early mucosal adenocarcinoma), AI did not detect a lesion, despite the presence of BERN (false negatives). One case of LGD was not recognized by any expert endoscopist;

► **Table 2** Performance of Barrett's nonexpert endoscopists with and without artificial intelligence assistance during the evaluation of video cases with nondysplastic Barrett's or Barrett's-related neoplasia. Group A: first view without AI, second view with AI; group B: first view with AI, second view without AI.

Healthcare setting	Examination without AI, % (95%CI)	Examination with AI, % (95%CI)	P value
<b>Overall (N = 16)</b>			
Group A			
▪ Sensitivity	69.8 (65.2–74.2)	78.0 (74.0–82.0)	0.001
▪ Specificity	67.3 (62.5–72.2)	72.7 (68.2–77.3)	0.01
▪ Accuracy	68.6 (65.3–71.9)	75.5 (72.5–78.5)	0.20
Group B			
▪ Sensitivity	73.1 (68.8–77.4)	73.1 (68.8–77.4)	>0.99
▪ Specificity	60.3 (55.2–65.2)	61.1 (56.2–66.3)	0.58
▪ Accuracy	67.1 (63.8–70.4)	67.5 (64.2–70.8)	0.74
<b>Secondary care hospitals (N = 4)</b>			
Group A			
▪ Sensitivity	68.0 (59.0–77.0)	72.0 (63.0–81.0)	0.42
▪ Specificity	63.6 (53.4–73.9)	75.0 (65.9–84.1)	0.05
▪ Accuracy	66.0 (59.0–72.9)	73.4 (67.0–79.8)	0.41
Group B			
▪ Sensitivity	73.1 (64.4–81.7)	73.1 (64.4–81.7)	>0.99
▪ Specificity	60.9 (51.1–70.7)	58.7 (48.9–68.5)	0.69
▪ Accuracy	67.3 (60.7–74.0)	66.3 (59.7–73.0)	0.79
<b>Private practice (N = 6)</b>			
Group A			
▪ Sensitivity	62.0 (54.0–69.3)	74.7 (67.3–81.3)	<0.001
▪ Specificity	74.2 (66.7–81.8)	75.8 (68.2–83.3)	0.77
▪ Accuracy	67.7 (62.1–73.0)	75.2 (70.2–80.1)	0.003
Group B			
▪ Sensitivity	66.0 (58.3–73.1)	67.9 (60.3–75.0)	0.45
▪ Specificity	63.8 (55.8–71.7)	65.2 (57.2–73.2)	0.50
▪ Accuracy	65.0 (59.5–70.4)	66.7 (61.2–72.1)	>0.99
<b>BE referral centers (N = 6)</b>			
Group A			
▪ Sensitivity	78.7 (72.0–85.3)	85.3 (79.3–90.7)	0.02
▪ Specificity	62.9 (54.5–71.2)	68.2 (59.8–75.8)	0.19
▪ Accuracy	71.3 (66.0–76.6)	77.3 (72.3–82.3)	0.74
Group B			
▪ Sensitivity	80.1 (73.7–85.9)	78.2 (71.2–84.6)	0.45
▪ Specificity	56.5 (48.6–64.5)	58.7 (50.7–66.7)	0.38
▪ Accuracy	69.1 (63.6–74.5)	69.0 (63.6–74.1)	0.15

AI, artificial intelligence; BE, Barrett's esophagus.

**Table 3** Stand-alone performance of an artificial intelligence system in the evaluation of Barrett's esophagus.

	Stand-alone performance of AI, % (95%CI)
Sensitivity	92.2 (88.2–95.6)
Specificity	68.9 (62.2–75.6)
Accuracy	81.3 (77.3–85.2)

AI, artificial intelligence-based clinical decision support system.

one case of HGD was not recognized by 3/6 expert endoscopists, and two further cases of mucosal cancer were not recognized by 2/6 and 3/6 endoscopists, respectively.

### Benchmarking tests with expert endoscopists

Expert endoscopists had an overall sensitivity of 83.3% (95%CI 79.1%–87.5%) without the support of AI and 85.0% (95%CI 81.0%–89.0%) with AI. Furthermore, their specificity was 58.1% (95%CI 52.2%–64.0%) and 58.9% (95%CI 53.0%–64.8%) without and with AI support, respectively. The overall accuracy of expert endoscopists in this trial was 71.5% (95%CI 67.8%–75.2%) without and 72.7% (95%CI 69.1%–76.4%) with the support of AI. There was no difference between group A and group B for expert endoscopists.

### Comparison of AI-assisted nonexperts with BE experts

Nonexpert endoscopists improved their performance significantly when using AI. However, the sensitivity of expert endoscopists without AI on this test set was still significantly superior to nonexpert performance with AI (83.3% [95%CI 79.1%–87.5%] vs. 75.5% [95%CI 72.4%–78.6%];  $P=0.005$ ). When comparing the specificity of nonexperts with the help of AI to experienced endoscopists without AI, we observed that nonexperts performed significantly better (58.1% [95%CI 52.2%–64.0%] vs. 66.8% [95%CI 64.4%–69.2%];  $P=0.01$ ). In terms of accuracy, experts without the support of AI were not superior to nonexperts with AI (71.5% [95%CI 67.8%–75.2%] vs. 71.4% [95%CI 69.1%–73.7%];  $P=0.96$ ).

### Influence of AI on the diagnostic confidence

With AI, participants indicated “low confidence” in 29.5% (95%CI 27.6%–31.4%) of video cases compared with 36.8% (95%CI 33.9%–39.7%) without AI, respectively. In 70.5% (95%CI 68.6%–72.4%) of video cases, participants indicated “high confidence” when using AI compared with 63.2% (95%CI 60.3%–66.1%;  $P<0.001$ ) of video case assessments without AI. Participants in groups A and B decided significantly more often with “high confidence” when using AI (group A:  $\Delta 8.5\%$  [95%CI 8.3%–8.7%]  $P<0.001$ ; group B:  $\Delta 6.2\%$  [95%CI 6.1%–6.3%]  $P<0.002$ ).

Irrespective of the order of appearance, when deciding with “high confidence,” all nonexperts (private practices, secondary care hospitals, BE referral centers, respectively) showed significantly better specificity than when deciding with “low confi-

dence” (81.7% [95%CI 77.9%–85.5%] vs. 38.0% [95%CI 30.8%–45.2%]  $P<0.001$ ; 90.0% [95%CI 85.4%–94.6%] vs. 40.0% [95%CI 33.5%–46.5%]  $P<0.001$ ; 72.7% [95%CI 68.1%–77.3%] vs. 40.4% [95%CI 33.7%–47.1%]  $P<0.001$ ).

Similarly, this effect of improved specificity in the three healthcare settings could be observed when using AI (79.4% [95%CI 75.5%–83.3%] vs. 47.4% [95%CI 39.9%–54.9%]  $P<0.001$ ; 89.4% [95%CI 85.1%–93.7%] vs. 41.9% [95%CI 34.8%–49.0%]  $P<0.001$ ; 72.8% [95%CI 68.6%–77.0%] vs. 38.7% [95%CI 30.5%–46.9%]  $P<0.001$ ). Overall, when using AI, participants decided more often with “high confidence.”

## Discussion

In this tandem, video-based trial, we found that nonexperts detected a higher proportion of Barrett's neoplasms when using AI. The effect of AI on performance was particularly prominent when AI was used during the second view than when videos were viewed immediately with AI.

The ESGE recommends that the performance of nonexperts in combination with AI should be comparable to that of expert endoscopists without AI [15]. In this trial, although nonexperts improved their sensitivity with AI, the sensitivity of expert diagnoses remained significantly higher than that of nonexperts. On the other hand, the diagnoses by nonexperts with AI were significantly more specific than those of experts. Subsequently, the overall diagnostic accuracy of nonexperts with AI was comparable to that of expert endoscopists without AI. In a similarly designed randomized, controlled tandem trial for gastric cancer lesions, Wu et al. demonstrated a significantly lower miss rate with AI and a significant improvement in cancer detection when AI was used in the second pass [25].

As described above, our tandem model study design [26, 27] included two groups (A and B), and showed that the use of AI after the conventional evaluation of BE videos by the human eye (group A) led to a significant improvement in the performance of the nonexperts. On the other hand, when AI was used directly and without initial human-eye evaluation (group B), no additional improvement was observed during the second view without AI. When correlating the influence of AI to the practice setting, we observed that physicians in private practice particularly benefitted from additional AI support. However, not all participants benefitted from AI equally. It remains unclear which factors influenced endoscopists' decisions to either follow or discard AI advice. Our current study suggests that human factors and human-computer interactions are of major importance in the context of AI and its applications.

Former AI trials on BE have usually compared the stand-alone performance of AI with the performance of endoscopists [28, 29]. However, the stand-alone performance is only a small fraction of the equation because endoscopists may or may not follow the suggestions of AI. Fockens et al. compared AI performance with that of endoscopists and described, depending on the test set, a sensitivity of between 88% and 100% during an image-based study [28]. Abdelrahim et al. demonstrated a sensitivity of more than 90% during a video-based study with 75 videos [29]. Both tests were limited to high definition

white-light endoscopy and reported only the stand-alone performance, without taking AI as a clinical decision support system into account. Furthermore, although net architectures that allow semantic segmentation were implemented, object detection with bounding boxes only were demonstrated. In the current study, AI allowed a multi-modal pixel-accurate segmentation of BERN in high definition white-light endoscopy, narrow-band imaging, and texture and color enhancement imaging, with continuous real-time CADe and CADx.

To better understand the decision-making process of endoscopists, we investigated how AI affects the level of confidence. Comparable to the effect that is observed when a more senior physician confirms a clinical decision of a less experienced physician, the diagnostic confidence of all BE nonexperts improved significantly with AI and was associated with better performance. However, diagnostic confidence is only one aspect of the human-machine interaction. Usability and user experience are further relevant factors to consider in future studies.

The development of AI in the field of BE remains challenging. Early BE lesions are subtle and difficult to discern, and determination of the histological or expert-opinion-based ground truth is challenging. In addition, the paucity of data for BE and BERN makes the training process of AI more difficult than, for example, for colonic colorectal polyps. Current commercially available AI systems for the colon provide bounding boxes for object detection and ROI demonstration to the user. Contrary to previously published trials in BE, where bounding boxes were used for object detection, we were able to implement a real-time pixel-precise delineation and segmentation of the ROI. This is particularly relevant as BE diagnosis and treatment involve detection and precise delineation of the ROI to improve targeted biopsy precision and pretherapeutic border recognition.

There are relevant study limitations to our video trial. First, our tandem study design may have introduced a possible bias because endoscopists assessing the video cases always saw each video case twice, without a “washout” time in between the assessments. A classical randomized controlled trial directly comparing two separate groups of endoscopists, one with and the other without AI, may have been better suited to assess the effect of AI on the performance of endoscopists because of the lower risk of bias. Second, BERN is often not limited to one single location but is multifocal. Although we had histological confirmation of the demonstrated lesions, sampling errors or false negatives are still possible. Furthermore, although we created a heterogeneous test set, including low grade inflammation and different levels of dysplasia, the final proportion of BERN lesions in the test set does not represent the true prevalence that endoscopists encounter in a real-world setting. Third, although 22 endoscopists participated in the trial, this sample size is considered relatively small, potentially limiting the generalizability of the findings, particularly concerning the subgroup analyses. Furthermore, a more positive attitude of the 22 participating endoscopists toward AI compared with the 11 endoscopists who were invited but did not participate in the trial may be a potential source of bias.

Regarding video case selection, we used high definition videos from a single center. In addition, this does not represent a true test of reality because AI should undergo evaluation with as much external data as possible. Finally, as we included 96 video cases from 72 patients, there may be a possibility of statistical dependency between the cases. However, video cases were chosen carefully to avoid visual overlaps between the video cases that were taken more than once from the same patient.

In conclusion, we developed and benchmarked an AI system to evaluate BE in standardized endoscopy videos. The stand-alone performance of AI was comparable to that of Barrett's experts. AI was especially beneficial to BE nonexpert endoscopists. Nonexpert endoscopists with the support of AI performed significantly better than without AI. AI seemed to confirm endoscopists' diagnoses while evaluating BE video cases, and higher diagnostic confidence appeared to correlate with improved performance. Further studies are needed to assess the effects of AI in clinical practice and better understand the various aspects of the human-computer interaction.

## Conflict of Interest

N. Parsa is VP of medical affairs at Satisfai Health. M. Byrne is CEO and Founder of Satisfai Health. H. Messmann has received lecture fees from Olympus, Ambu, IPSEN, Medtronic, and Falk, and research grants from Olympus and Satisfai; he is also a consultant for Satisfai. A. Ebigbo has held lectures for Olympus, Fuji, Pentax, Medtronic, Falk, and Ambu. The remaining authors declare that they have no conflict of interest.

## References

- [1] Hvid-Jensen F, Pedersen L, Drewes AM et al. Incidence of adenocarcinoma among patients with Barrett's esophagus. *N Engl J Med* 2011; 365: 1375–1383
- [2] Coleman HG, Xie SH, Lagergren J. The epidemiology of esophageal adenocarcinoma. *Gastroenterology* 2018; 154: 390–405 doi:10.1053/j.gastro.2017.07.046
- [3] Sung H, Ferlay J, Siegel RL et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021; 71: 209–249 doi:10.3322/caac.21660
- [4] Smyth EC, Lagergren J, Fitzgerald RC et al. Oesophageal cancer. *Nat Rev Dis Primers* 2017; 3: 17048 doi:10.1038/nrdp.2017.48
- [5] Visrodia K, Singh S, Krishnamoorthi R et al. Magnitude of missed esophageal adenocarcinoma after Barrett's esophagus diagnosis: a systematic review and meta-analysis. *Gastroenterology* 2016; 150: 599–607.e597
- [6] Messmann H, Ebigbo A, Hassan C et al. How to integrate artificial intelligence in gastrointestinal practice. *Gastroenterology* 2022; 162: 1583–1586 doi:10.1053/j.gastro.2022.02.029
- [7] van der Sommen F, Zinger S, Curvers WL et al. Computer-aided detection of early neoplastic lesions in Barrett's esophagus. *Endoscopy* 2016; 48: 617–624 doi:10.1055/s-0042-105284



- [8] de Groof AJ, Struyvenberg MR, van der Putten J et al. Deep-learning system detects neoplasia in patients with Barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking. *Gastroenterology* 2020; 158: 915–929.e914
- [9] de Groof AJ, Struyvenberg MR, Fockens KN et al. Deep learning algorithm detection of Barrett's neoplasia with high accuracy during live endoscopic procedures: a pilot study (with video). *Gastrointest Endosc* 2020; 91: 1242–1250
- [10] Hashimoto R, Requa J, Dao T et al. Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in Barrett's esophagus (with video). *Gastrointest Endosc* 2020; 91: 1264–1271.e1261
- [11] Iwagami H, Ishihara R, Aoyama K et al. Artificial intelligence for the detection of esophageal and esophagogastric junctional adenocarcinoma. *J Gastroenterol Hepatol* 2021; 36: 131–136 doi:10.1111/jgh.15136
- [12] Struyvenberg MR, de Groof AJ, van der Putten J et al. A computer-assisted algorithm for narrow-band imaging-based tissue characterization in Barrett's esophagus. *Gastrointest Endosc* 2021; 93: 89–98 doi:10.1016/j.gie.2020.05.050
- [13] Hussein M, González-Bueno Puyal J, Lines D et al. A new artificial intelligence system successfully detects and localises early neoplasia in Barrett's esophagus by using convolutional neural networks. *United European Gastroenterol J* 2022; 10: 528–537
- [14] Ebigo A, Mendel R, Probst A et al. Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. *Gut* 2019; 68: 1143–1145 doi:10.1136/gutjnl-2018-317573
- [15] Messmann H, Bisschops R, Antonelli G et al. Expected value of artificial intelligence in gastrointestinal endoscopy: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. *Endoscopy* 2022; 54: 1211–1231 doi:10.1055/a-1950-5694
- [16] Vasey B, Nagendran M, Campbell B et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med* 2022; 28: 924–933 doi:10.1038/s41591-022-01772-9
- [17] Chen L-C, Zhu Y, Papandreou G et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C et al. (eds.) *Computer Vision – ECCV 2018*. Cham: Springer International Publishing; 2018: 833–851
- [18] Huang Y, Wang Q, Jia W et al. See more than once: kernel-sharing atrous convolution for semantic segmentation. *Neurocomputing* 2021; 443: 26–34
- [19] He K, Zhang X, Ren S et al. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016: 770–778
- [20] Mendel R, Rauber D, de Souza LA Jr et al. Error-Correcting Mean-Teacher: corrections instead of consistency-targets applied to semi-supervised medical image segmentation. *Comput Biol Med* 2023; 154: 106585 doi:10.1016/j.compbimed.2023.106585
- [21] Singer ME, Odze RD. High rate of missed Barrett's esophagus when screening with forceps biopsies. *Esophagus* 2023; 20: 143–149 doi:10.1007/s10388-022-00943-4
- [22] Ebigo A, Mendel R, Probst A et al. Real-time use of artificial intelligence in the evaluation of cancer in Barrett's oesophagus. *Gut* 2020; 69: 615–616 doi:10.1136/gutjnl-2019-319460
- [23] Meinikheim M, Mendel R, Scheppach MW et al. Influence of an artificial intelligence (AI) based decision support system (DSS) on the diagnostic performance of non-experts in Barrett's esophagus related neoplasia (BERN). *Endoscopy* 2022; 54: OP076
- [24] Weusten B, Bisschops R, Coron E et al. Endoscopic management of Barrett's esophagus: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. *Endoscopy* 2017; 49: 191–198 doi:10.1055/s-0042-122140
- [25] Wu L, Shang R, Sharma P et al. Effect of a deep learning-based system on the miss rate of gastric neoplasms during upper gastrointestinal endoscopy: a single-centre, tandem, randomised controlled trial. *Lancet Gastroenterol Hepatol* 2021; 6: 700–708
- [26] Glissen Brown JR, Mansour NM, Wang P et al. Deep learning computer-aided polyp detection reduces adenoma miss rate: a United States multi-center randomized tandem colonoscopy study (CADET-CS Trial). *Clin Gastroenterol Hepatol* 2022; 20: 1499–1507.e1494
- [27] Wallace MB, Sharma P, Bhandari P et al. Impact of artificial intelligence on miss rate of colorectal neoplasia. *Gastroenterology* 2022; 163: 295–304.e295
- [28] Fockens KN, Jukema JB, Boers T et al. Towards a robust and compact deep learning system for primary detection of early Barrett's neoplasia: initial image-based results of training on a multi-center retrospectively collected data set. *United European Gastroenterol J* 2023; 11: 324–336 doi:10.1002/ueg2.12363
- [29] Abdelrahim M, Saiko M, Maeda N et al. Development and validation of artificial neural networks model for detection of Barrett's neoplasia: a multicenter pragmatic nonrandomized trial (with video). *Gastrointest Endosc* 2023; 97: 422–434