Accepted Manuscript

Submission Date: 2023-09-19
Accepted Date: 2024-03-13
Accepted Manuscript online: 2024-03-28

Endoscopy

Effect of AI on performance of endoscopists to detect Barrett neoplasia: A Randomized Tandem Trial

Michael Meinikheim, Robert Mendel, Christoph Palm, Andreas Probst, Anna Muzalyova, Markus W Scheppach, Sandra Nagl, Elisabeth Schnoy, Christoph Römmele, Dominik A Schulz, Jakob Schlottmann, Friederike Prinz, David Rauber, Tobias Rückert, Tomoaki Matsumura, Glòria Fernández-Esparrach, Nasim Parsa, Michael F Byrne, Helmut Messmann, Alanna Ebigbo.

Affiliations below.

DOI: 10.1055/a-2296-5696

Please cite this article as: Meinikheim M, Mendel R, Palm C et al. Effect of AI on performance of endoscopists to detect Barrett neoplasia: A Randomized Tandem Trial. Endoscopy 2024. doi: 10.1055/a-2296-5696

Conflict of Interest: -Nasim Parsa is VP of medical affairs at Satisfai Health.

-Michael Byrne is CEO and Founder of Satisfai Health.

-Helmut Messmann received lecture fees from Olympus, Ambu, IPSEN, medtroninc, Falk and received research grants from Olmypus and Satisfai. Furthermore, he is a Satisafi consultant.

-Alanna Ebigbo has held lectures for Olympus, Fuji, Pentax, Medtronic, Falk and Ambu

The remaining Authors declare that there is no conflict of interest.

Abstract:

Background and study aims

To evaluate the effect of an AI-based clinical decision support system (AI) on the performance and diagnostic confidence of endoscopists during the assessment of Barrett's esophagus (BE).

Patients and Methods

Ninety-six standardized endoscopy videos were assessed by 22 endoscopists from 12 different centers with varying degrees of BE experience.

The 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 standalone sensitivity, specificity, and accuracy of 92.2%, 68.9%, and 81.6%, 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 69.7% (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.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Thieme

This article is protected by copyright. All rights reserved.

Accepted Manuscript

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

Corresponding Author:
Dr. Michael Meinikheim, University Hospital Augsburg, Department of Gastroenterology, Augsburg, Germany, michael.meinikheim@med.uni-augsburg.de

Affiliations:
Michael Meinikheim, University Hospital Augsburg, Department of Gastroenterology, Augsburg, Germany
Robert Mendel, Ostbayerische Technische Hochschule Regensburg (OTH Regensburg), Regensburg Medical Image Computing (ReMIC), Regensburg, Germany
Robert Mendel, Ostbayerische Technische Hochschule Regensburg (OTH Regensburg), Regensburg Center of Health Sciences and Technology, Regensburg, Germany
[...]
Alanna Ebigbo, University Hospital Augsburg, Department of Gastroenterology, Augsburg, Germany

This article is protected by copyright. All rights reserved.
Accepted Manuscript

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Effect of AI on performance of endoscopists to detect Barrett neoplasia: A Randomized Tandem Trial

Michael Meinikheim*[1], Robert Mendel*[2], Christoph Palm[2], Andreas Probst[1], Anna Muzalyova[1], Markus Scheppach[1], Sandra Nagl[1], Elisabeth Schnoy[1], Christoph Römmele[1], Dominik Schulz[1], Jakob Schlottmann[1], Friederike Prinz[1], David Rauber[2], Tobias Rückert[2], Tomoaki Matsumura[3], Gloria Fernandez Esparrach[4,5,6,7], Nasim Parsa[8,10], Michael Byrne[9,10], Helmut Messmann**[1], Alanna Ebigbo**[1]

*contributed equally
** shared last authorship

Institute
1 University Hospital Augsburg, Department of Gastroenterology, Augsburg, Germany
2 Ostbayerische Technische Hochschule Regensburg, Regensburg Medical Image Computing, Regensburg, Germany
3 Department of Gastroenterology, Chiba University Graduate School of Medicine 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, AZ, USA
9 Vancouver General Hospital, The University of British Columbia, Vancouver, British Columbia, Canada
10 Satisfai Health, Vancouver, British Columbia, Canada

Contact corresponding author:
Michael Meinikheim, MD; e-mail: Michael.meinikheim@uk-augsburg.de; phone: 0821 400 2351

Authors contributions:

Michael Meinikheim: Contributed to the development of the AI-system, contributed to the development of analysis tools, conceived and designed the analysis, collected the data, performed the analysis and drafted the paper. Furthermore, he was responsible for project administration
Robert Mendel: Contributed to the development of the AI-system, contributed analysis tools, conceived and designed the analysis, performed the analysis and drafted the paper. Furthermore, he was responsible for project administration. Robert Mendl contributed equally as Michael Meinikheim to this manuscript.
Christoph Palm: Contributed to the development of the AI-system, contributed analysis tools, conceived and designed the analysis, performed the analysis and drafted the paper. Furthermore, he was responsible for funding acquisition, supervision and validation.
Andreas Probst: Contributed to the development of the AI-system, collected the data and drafted the paper.
Anna Muzalyova: Conceived and designed the analysis, performed the analysis, contributed analysis tools and drafted the paper.
Markus Scheppach: Contributed to the development of the AI-system, collected the data and drafted the paper.
Sandra Nagl: Collected the data, contributed to the development of the AI-system and drafted the paper.
Elisabeth Schnoy: Collected the data, contributed to the development of the AI-system and drafted the paper.
Christoph Römmele: Collected the data, contributed to the development of the AI-system and drafted the paper.
Dominik Schulz: Collected the data, contributed to the development of the AI-system and drafted the paper.
Jakob Schlottmann: Collected the data, contributed to the development of the AI-system and drafted the paper.
Friederike Prinz: Collected the data, contributed to the development of the AI-system and drafted the paper.
David Rauber: Contributed to the development of the AI-system, contributed analysis tools and drafted the paper.
Tobias Rückert: Contributed to the development of the AI-system, contributed analysis tools and drafted the paper.
Tomoaki Matsumura: Contributed data, contributed to the development of the AI-system and drafted the paper.
Gloria Fernandez Esparrach: Contributed data, contributed to the development of the AI-system and drafted the paper.
Nasim Parsa: Contributed to the development of the AI-system and drafted the paper
Michael Byrne: Contributed to the development of the AI-system and drafted the paper
Helmut Messmann: Contributed to the development of the AI-system, conceived and designed the analysis, collected the data, performed the analysis and drafted the paper. Furthermore, he was responsible for funding acquisition, supervision and validation.
Alanna Ebigbo: Contributed to the development of the AI-system, conceived and designed the analysis, collected the data, performed the analysis and drafted the paper. Furthermore, he was responsible for funding acquisition, supervision and validation.

This article is protected by copyright. All rights reserved.
Accepted Manuscript

ABSTRACT

Background and study aims

To evaluate the effect of an AI-based clinical decision support system (AI) on the performance and diagnostic confidence of endoscopists during the assessment of Barrett's esophagus (BE).

Patients and Methods

Ninety-six standardized endoscopy videos were assessed by 22 endoscopists from 12 different centers with varying degrees of BE experience.

The 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 standalone sensitivity, specificity, and accuracy of 92.2%, 68.9%, and 81.6%, 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 69.7% (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 the additional AI. BE experts and nonexperts remained significantly below the standalone performance of AI, suggesting that there may be other factors influencing endoscopists to follow or discard AI advice.

This article is protected by copyright. All rights reserved.

Accepted Manuscript

Abbreviations

AI – Artificial Intelligence
BE - Barrett's esophagus
BERN - Barrett's esophagus-related neoplasia
CADe - Computer-aided detection
CADx - Computer-aided diagnosis
ESGE - European Society of Gastrointestinal Endoscopy
HD-WLE - High-definition White Light Endoscopy
HGD - High-grade dysplasia
LGD - Low-grade dysplasia
NBI - Narrow Band Imaging
NDBE - Nondysplastic Barrett's esophagus
ROI - Region of interest
TXI - Texture and Color Enhancement Imaging

This article is protected by copyright. All rights reserved.
Accepted Manuscript

Introduction

Barrett's esophagus (BE) is a precursor of esophageal adenocarcinoma. Even though studies suggest that the rate of progression of non-dysplastic 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 demonstrate 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 has undergone intense research in endoscopy with numerous potential applications[6]. One possibility of Artificial Intelligence is to offer a "second opinion" or decision support during the endoscopic evaluation of BE. Several research teams have used deep learning to develop artificial intelligence-based clinical decision support systems (AI) for computer-aided detection (CADe) and computer-aided diagnosis (CADx) in the context of BE assessment and BERN [7-13]. 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 standalone performance of AI and compared it to the standalone 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 has on the performance of BE nonexpert endoscopists assessing a Barrett's video data set.

This article is protected by copyright. All rights reserved.

Accepted Manuscript

This article is protected by copyright. All rights reserved.

Accepted Manuscript

Material & Methods

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

Study Outcomes

Primary outcome:

- Effect of AI on the diagnostic performance of nonexpert endoscopists in BE evaluation.

Secondary outcomes:

- Standalone performance of AI for the detection and segmentation of BERN.
- Effect of AI on the diagnostic performance of expert endoscopists in BE evaluation.
- 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 (HD-WLE), Narrow Band Imaging (NBI), Texture and Color Enhancement Imaging (TXI) 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 as “at risk” show histologically confirmed BERN, from a distance or perspective that does not allow an accurate visual assessment. More detailed descriptions are attached in the supplementary files (S1).

Deep learning Model

The deep learning model is based on the DeepLabV3+ [17] architecture with Kernel-Sharing [18] and a ResNet50 [19] backbone.
The segmentation task is trained with the semi-supervised ECMT [20] algorithm. More detailed descriptions are attached in the supplementary files (S2).

Algorithm

The algorithm integrates information into the trained model to provide consistent predictions. **Figure 1** offers 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 models predictions marked on the screen. More detailed descriptions are attached in the supplementary files (S3).

Description of the video trial data

Ninety-six (96) prospectively collected videos of endoscopic examinations in 72 consecutive patients who presented to the University Hospital of Augsburg to evaluate BE and BERN between the 1st of October 2021 and 30th of September 2022 and fitting the study criteria were included. Patients included 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 length between 15 seconds and 90 seconds. While most videos showed the entire segment of the BE, 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 included in this trial were exclusively flat or slightly elevated (Paris IIa/IIb) lesions **(Table 1)**.
All included cases contained at least two imaging modalities, including HD-WLE, NBI, or TXI. Data were obtained from endoscopic examinations with Olympus GIF-HQ190, GIF-XZ1200, GIF-EZ1500 gastroscopes, and CV-1500 Evis X1 endoscopic processor. Video documentation of forceps biopsy or endoscopic resection of 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 so that there was no visual overlap 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 secondly with additional AI (group A). The other half of the cases were presented in the opposite order, first with and secondly without additional AI support (group B). In addition, the data set 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 (**Figure 2**).

Evaluation of the influence of AI on the 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” from 0 – 4 and “high confidence” from 5 – 9 regarding how sure or unsure participants were of their diagnoses.

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

This article is protected by copyright. All rights reserved.
Accepted Manuscript

was expert assessment which was confirmed histologically. Specificity was defined as the correct diagnosis of video cases without a visible neoplasm as nondysplastic BE (NDBE).
Based on previous work [14, 21-23], 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 and fitting the inclusion criteria during the period from October 1st 2021 until September 30th 2022. As described above, 96 video cases were then generated from these 72 patients included, 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. Since 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 with McNemar's test. We used Wald interval as method to determine the confidence intervals. The performance of nonexpert endoscopists with additional AI was compared to the benchmark performance of Barrett's experts and tested for statistically significant differences with the Chi-Square test. Also, differences in performance depending on the confidence level were tested using Chi-Square Test. The significance level was set at 0.05. All statistical tests were performed using SPSS Version 28.0.

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 attached in supplementary table (ST1). BE experts were defined according to the position statement of the ESGE including endoscopists with regular BE evaluation and with an 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, 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 (Supplementary, S4). 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 re-assessed 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 non-stable predictions to evaluate the persistency of a prediction by AI. Stable prediction was defined as a segmentation heat map displayed for more than three seconds (150 consecutive frames). Non-stable prediction implied cases where the segmentation map repeatedly appeared at the same spot for an overall cumulative time of more than three seconds (150 frames) but not continuously **(Figure 3)**.

This article is protected by copyright. All rights reserved.
Accepted Manuscript

This article is protected by copyright. All rights reserved.

Accepted Manuscript

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 69.7% (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 and SF 1 in the supplementary material**).

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 did not change after the first review with AI **(Table 2).**

Standalone 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 NDBE (specificity of 68.9% (95% CI, 62.2% - 75.6%). The system's overall accuracy on this test set was 81.6% (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 ≥ 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 less than 150 frames on the respective lesion (non-stable prediction). False-positive results appeared in 14 cases (six non-stable predictions with false-positive segmental overlays of less than 150 frames and eight stable predictions with false-positive segmental overlays of more than 150 frames). In four cases (1x LGD, 1x HGD, 2x 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; 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) **(Table 3)**.

Benchmarking tests with expert endoscopists

Expert endoscopists had an overall sensitivity of 83.3% (CI 95%, 79.1%-87.5%) without the support of AI and 85.0% (CI 95%, 81.0%-89.0%) with AI. Furthermore, their specificity was 58.1% (CI 95%, 52.2%-64.0%) and 58.9% (CI 95%, 53.0%-64.8%) without and with AI support. The overall accuracy of

expert endoscopists in this trial was 71.5% (CI 95%, 67.8% - 75.2%) without and 72.7% (CI 95%, 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 to BE experts

Nonexpert endoscopists improved their performance significantly when using AI. However, the sensitivity of expert endoscopists on this test set was still significantly superior to nonexpert performance with AI (83.3% (CI 95%, 79.1%-87.5%) vs. 75.5% (CI 95%, 72.4%-78.6%); p = 0.005). When comparing the specificity of nonexperts with the help of AI to experienced endoscopists, we observed that nonexperts performed significantly better (58.1% (CI 95%, 52.2%-64.0%) vs. 66.8% (CI 95%, 64,4% - 69,2%); p = 0.011). In terms of accuracy, experts without the support of AI were not superior to nonexperts with AI (71.5% (CI 95%, 67.8% - 75.2%) vs. 71.4% (CI 95%, 69.1%-73.7%; p = 0.961).

Influence of AI on the diagnostic confidence

With AI, participants indicated “low confidence” in 29.5% (CI 95%, 27.6% - 31.4%) of video cases compared with 36.8% (CI 95%, 33.9% - 39.7%) without AI, respectively. In 70.5% (CI 95%, 68.6% - 72.4%) of video cases, participants indicated “high confidence” when using AI compared to 63.2% (CI 95%, 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 (Δ8.5% (CI 95%, 8.3% - 8.7%); (p<0.001) and Δ6.2% (CI 95%, 6.1% - 6.3); (p<0.002).

Irrespective of the order of appearance, when deciding with “high confidence”, all nonexpert (private practices, secondary care hospitals, BE referral centers, respectively) showed significantly better specificity compared to when deciding with “low confidence” (81.7% (CI 95% 77.9% - 85.5%) vs. 38.0% (CI 95%, 30.8% - 45.2%), P < 0.001; 90.0% (CI 95%, 85.4% - 94.6%) vs. 40.0% (CI 95%, 33.5% - 46.5%), P<0.001; 72.7.% (CI 95%, 68.1% - 77.3% vs. 40.4.% (CI 95%, 33.7% - 47.1%, P< 0.001).
Likewise, this effect could be observed when using AI (79.4% (CI 95%, 75.5% - 83.3%) vs. 47.4% (CI 95%, 39.9% - 54.9%, P<0.001; 89.4% (CI 95% 85.1% - 93.7%) vs. 41.9% (CI 95% 34.8% - 49.0%), P<0.001; 72.8% (CI 95% 68.6% - 77.0%) vs. 38.7% (CI 95%, 30.5% - 46.9%), P<0.001). Overall, when using AI, participants decided more often with “high confidence”.

This article is protected by copyright. All rights reserved.
Accepted Manuscript

This article is protected by copyright. All rights reserved.

Accepted Manuscript

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 on second view than when videos were viewed immediately with AI.

The ESGE recommends that nonexperts' performance in combination with AI should be comparable with expert endoscopists[15]. In this trial, even though nonexperts improved their sensitivity, experts remained significantly more sensitive than nonexperts. On the other hand, nonexperts with AI were significantly more specific. Subsequently, the overall diagnostic accuracy of nonexperts with AI was comparable with 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 of 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, no additional improvement was observed (group B). When correlating the influence of AI to the area of practice, we observed that particularly physicians in the private practice group benefitted from the additional AI. However, not all participants benefitted from AI equally. It remains unclear which factors influenced endoscopists 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 standalone performance of AI with the performance of endoscopists [28, 29]. However, the standalone 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 to endoscopists and described, depending on the test set, a sensitivity 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 HD-WLE and reported only the standalone performance without taking AI as a clinical decision support system into account. Furthermore, even though net architectures that allow semantic segmentation were implemented, object detection with bounding boxes only were demonstrated. In this study, AI allowed a multi-modal pixel-accurate segmentation of BERN in HD-WLE, NBI, and TXI 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 reconfirms 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 the determination of the histological or expert-opinion-based ground truth is challenging. Also, 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 in 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 ROI. This is particularly relevant since BE diagnosis and treatment involve detection and precise delineation of the ROI to improve targeted biopsy precision and pre-therapeutic border recognition.

There are relevant study limitations in our video trial. Firstly, our tandem study design may have introduced a possible bias because endoscopist 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 group 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. Secondly, BERN is often not limited to one single location but is multifocal. Even though we had histological confirmation of the demonstrated lesions, sampling errors or false negatives are still possible. Furthermore, although we created a heterogenous 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 in a real-world setting will encounter. Thirdly, 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 towards AI compared to the 11 endoscopists who were addressed but didn´t participate in the trial may be a potential source of bias.

Regarding video case selection, we used high-definition videos from a single center. Also, this does not represent a true test of reality because AI should undergo evaluation with as much external data as possible. Finally, since 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 AI to evaluate BE in standardized endoscopy videos. The standalone 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 seemed to reconfirm endoscopists while evaluating BE video cases, and higher diagnostic confidence appears 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 human-computer interaction.

This article is protected by copyright. All rights reserved.

Accepted Manuscript

This article is protected by copyright. All rights reserved.

Accepted Manuscript

References:

1. Hvid-Jensen F, Pedersen L, Drewes AM et al. Incidence of Adenocarcinoma among Patients with Barrett's Esophagus. New England Journal of Medicine 2011; 365: 1375-1383. doi:10.1056/NEJMoa1103042
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; quiz e514-595. doi:10.1053/j.gastro.2015.11.040
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. doi:10.1053/j.gastro.2019.11.030
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. doi:10.1016/j.gie.2019.12.048
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. doi:10.1016/j.gie.2019.12.049
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. doi:10.1002/ueg2.12233
14. Ebigbo 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. Nature Medicine 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. doi:https://doi.org/10.1016/j.neucom.2021.02.091
19. He K, Zhang X, Ren S et al. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770-778. DOI: 10.1109/CVPR.2016.90
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.compbiomed.2023.106585
21. Singer ME, Odze RD. High rate of missed Barrett's esophagus when screening with forceps biopsies. Esophagus 2022. doi:10.1007/s10388-022-00943-4
22. Ebigbo 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. doi:10.1055/s-0042-1744639
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. doi:10.1016/s2468-1253(21)00216-8
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. doi:10.1016/j.cgh.2021.09.009
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. doi:10.1053/j.gastro.2022.03.007
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. doi:10.1016/j.gie.2022.10.031

Figure legends:

Figure 1 – Algorithm overview
Figure 2 – Schematic representation of the test set for the participants of this trial
Figure 3 – Included images in the written instructions for the participants of the trial
Table 1 - Distribution of histology and length of Barrett's-esophagus segment
Table 2 – Performance of Barrett non-experienced endoscopists with and without AI during the evaluation of video cases with non-dysplastic Barrett's or Barrett's related neoplasia.
Table 3 – Standalone performance of an AI system in the evaluation of Barrett's esophagus.

This article is protected by copyright. All rights reserved.
Accepted Manuscript

This article is protected by copyright. All rights reserved.

Accepted Manuscript

Supplementary material:

S1.
Training data

The dataset used to train and cross-validate the model contains labeled and unlabeled images, including samples recorded during endoscopic procedures and frames of video sequences. The complete dataset consisted of images from 557 patients, including 51273 images. A subset of this dataset, designated partially labeled, contained 48063 images. The partially labeled subset included images derived from videos of 97 patients. No pixel-level data is accessible in the partially labeled subset, but image-level information is available for a fraction of the images. These include 9046 NDBE images and 156 BERN images. This partially labeled subset originated from either videos of patients without BERN, or additional images of perspective shifts of areas also included in the fully labeled portion of the dataset. The model's capacity to distinguish between overview and close-up perspectives, and to label distant suspicious regions as "risk" while reserving the BERN label for close-up predictions, enables this partitioning approach. 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 conducted by BE expert endoscopists. The training dataset was diverse, encompassing multiple modalities, including high-definition white light endoscopy (HD-WLE), Narrow Band Imaging (NBI), and Texture and Color Enhancement Imaging (TXI). In addition, images of NDBE and BERN from acetic acid and indigo carmine chromoendoscopy were included. The training dataset included an overview and near-focus images of the region of interest (ROI) for a single lesion. These perspective shifts were encoded as binary labels and included in the 3210 images of the labeled subset.

S2.
Deep learning (DL) Model

The DL model builds upon an augmented DeepLabV3+ model with Kernel-Sharing in the atrous pooling block and a ResNet50 backbone. The model optimizes the solution for the four-class segmentation problem spanning the following categories: normal tissue, NDBE, BERN, and regions at risk. We included the label "regions at risk" to improve the representation of areas where a BERN is possible; however, a more detailed inspection of the respective region is required to confirm the diagnosis. For instance, a near-focus image of a histologically confirmed BERN case during segmentation would be assigned to BERN. In contrast, an image of a fraction of the same lesion from a long distance will not be classified as BERN with absolute certainty. Such images or frames received the label "region at risk." This approach prevents overloading the BERN segmentation task and forces the model to distinguish between differences due to varying perspectives, such as the distance between the endoscope and the ROI.
Moreover, this segmentation model extends with two auxiliary classification branches. The first auxiliary branch connects to the features of an early residual block of the resnet50 backbone. Each auxiliary classifier consists of two linear layers that increase the feature dimension, followed by Batch Normalization, ReLU activation, and dropout layers. The final linear layer projects the features to class dimensions. The first binary auxiliary classifier aims to differentiate between the perspective of the current image. It classifies images as either overview images of the esophagus or close-up and near-focus images. The second auxiliary classifier attaches to the pooled features of the segmentation-specific decoder of the model and follows the same architecture as the first auxiliary classifier. This auxiliary classifier optimizes the binary BERN detection problem. The pooled features were intended to add a global context to the multi-scale feature detection in the decoder. Our architecture design further strengthens this relationship by directly utilizing these features for the image-level BERN detection task and thus helps the model in the pixel-level task. The semi-supervised training approach follows the ECMT method. A secondary teacher model, derived as an exponential moving average of the weights of the primary model, is used to correct predicted mistakes in a given image and segmentation input pair. On unlabeled images, the combination of the predicted segmentation of the primary model and the corrections of the secondary model forms the pseudo labels that the model is optimized on. In addition to the training procedure and architectural considerations outlined in ECMT, we continue training the auxiliary BERN classifier on images without pixel-level but image-level information.

This article is protected by copyright. All rights reserved.
Accepted Manuscript

S3.
Algorithm

The core of our algorithm is the trained model, which is then extended with temporal information and buffered and stabilized predictions to deliver consistent and reliable results. Figures 1 and 2 offer a comprehensive overview of the components involved. For every incoming frame, the forward pass of the trained neural network is split into encoding and decoding steps. In parallel to the encoding step, we estimate the change between the previous and current input frame and utilize this resulting temporal information to adjust the buffered encoded information of the previous frame. This step enables the algorithm to include temporal information with accurate spatial information and utilize buffered encodings from previous iterations for the auxiliary classification and, more importantly, for the segmentation task. The resulting temporarily corrected predictions will influence and be subject to post-processing steps for each frame. Regions predicted as BERN but with low confidence will be reassigned with the "uncertain" label, not present during training. Furthermore, the binary perspective prediction result will influence marking the identified area as BERN or "at risk." Therefore, the size of the BERN region and both binary auxiliary tasks and the temporal information determine the rate of change of an internal stability counter. With the post-processed prediction, the stability determines when the algorithm will tag an input as BERN and functions as a preventive measure for false positives, which could disturb or distract potential users even when apparent only for a split second. Both stability and temporal adjustments help stabilize the model output and make it more robust to outliers. We further focus on this aspect by calculating metrics from the estimated change between the current and previous input image. Apart from the role of the metrics as part of the algorithm, they are also relayed to the user in the motion history graph (Figure 2). Large, fast, or chaotic movements, derived from the estimated temporal information, correlate with deteriorating image quality and will flag an input as unsuitable for our model. In these cases, we reset the stability calculation and the temporal buffer that adjusts and merges encoded features and allow both to replenish before our algorithm continues to display its prediction. This is indicated by subtle changes in the icons in the graphical user interface and the motion history graph.

S4.
Software trial tool

We developed a software tool to enable participants to assess video clips of BE cases. To simulate real-life examination situations, our software tool allowed video clips to be paused at any time, and participants could go back to previous frames, allowing a detailed inspection of the entire clip. Furthermore, the software included an annotation function to demarcate an ROI. Therefore, when participants assumed a BERN, the software firstly allowed them to delineate the area where BERN was suspected and, secondly, demarcate a spot where participants would take a biopsy to confirm their suspicion histologically. Before participants could proceed to the next case, they were required to indicate on a scale from 0-9 how confident they were in the correctness of their diagnosis; zero being extremely insecure, and nine being extremely confident about the correctness of their diagnosis. Before the trial, all participants received a study manual with written instructions and a video tutorial about the features of the AIS as well as the trial framework. After completing the instructions, all participants were required to participate in a brief training which included a test run on how to proceed when NDBE or BERN was suspected.

This article is protected by copyright. All rights reserved.
Accepted Manuscript

Supplementary table:

Participants	Gender	Age	Years of endoscopy experience
BA	M	>50	>20 years
BF	M	>50	>20 years
BH	F	30-40	10-20 years
BJ	M	30-40	10-20 years
EA	M	40-50	10-20 years
FD	M	30-40	<10 years
FEG	F	40-50	10-20 years
GD	F	40-50	10-20 years
GLF	M	30-40	<10 years
HRD	M	>50	>20 years
HS	F	40-50	10-20 years
KR	M	40-50	>20 years
MT	M	40-50	>20 years
NS	F	30-40	<10 years
OK	M	30-40	<10 years
PA	M	>50	>20 years
PF	F	30-40	<10 years
RC	M	30-40	<10 years
SA	F	40-50	10-20 years
SJ	M	30-40	<10 years
SM	M	30-40	<10 years
SM	M	30-40	<10 years

ST 1 – Participating endoscopists

This article is protected by copyright. All rights reserved.

Accepted Manuscript

Tables 1- 3:

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

Table 1 - Distribution of histology and length of Barrett's-esophagus segment
Abbreviations – NDBE: Non-dysplastic Barrett's esophagus; LGD: Low-grade dysplasia; HGD: High-grade dysplasia; T1a and T1b: According to TNM-Classification of malignant tumors

This article is protected by copyright. All rights reserved.

Accepted Manuscript

OVERALL		*Examination*	*Examination*	*p-*

			without AI (95% CI)	*with AI (95% CI)*	*value*
(N = 16)	**Group A (Routine-first)**	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%)	0.014
		Accuracy	68.6% (65.3% - 71.9%)	75.5% (72.5% - 78.5%)	0.202
	Group B (With AI first)	Sensitivity	73.1% (68.8% - 77.4%)	73.1% (68.8% - 77.4%)	1.000
		Specificity	60.3% (55.2% - 65.2%)	61.1% (56.2% - 66.3%)	0.581
		Accuracy	67.1% (63.8% - 70.4%)	67.5% (64.2% - 70.8%)	0.736
Endoscopists at secondary care hospitals (N = 4)	**Group A (Routine-first)**	Sensitivity	68.0% (59.0% - 77.0%)	72.0% (63.0% - 81.0%)	0.424
		Specificity	63.6% (53.4% - 73.9%)	75.0% (65.9% - 84.1%)	0.052
		Accuracy	66.0% (59.0% - 72.9%)	73.4% (67.0% - 79.8%)	0.405
	Group B (With AI first)	Sensitivity	73.1% (64.4% - 81.7%)	73.1% (64.4% - 81.7%)	1
		Specificity	60.9% (51.1% - 70.7%)	58.7% (48.9% - 68.5%)	0.687
		Accuracy	67.3% (60.7% - 74.0%)	66.3% (59.7% - 73.0%)	0.791
Endoscopists in private practice (N = 6)	**Group A (Routine-first)**	Sensitivity	62.0% (54.0% - 69.3%)	74.7% (67.3% - 81.3%)	0
		Specificity	74.2% (66.7% - 81.8%)	75.8% (68.2% - 83.3%)	0.774
		Accuracy	67.7% (62.1% - 73.0%)	75.2% (70.2% - 80.1%)	0.003
	Group B (With AI first)	Sensitivity	66.0% (58.3% - 73.1%)	67.9% (60.3% - 75.0%)	0.453
		Specificity	63.8% (55.8% - 71.7%)	65.2% (57.2% - 73.2%)	0.5
		Accuracy	65.0% (59.5% - 70.4%)	66.7% (61.2% - 72.1%)	1
Endoscopists at Barrett-referral centers (N = 6)	**Group A (Routine-first)**	Sensitivity	78.7% (72.0% - 85.3%)	85.3% (79.3% - 90.7)	0.021
		Specificity	62.9% (54.5% - 71.2%)	68.2% (59.8% - 75.8%)	0.189
		Accuracy	71.3% (66.0% - 76.6%)	77.3% (72.3% - 82.3%)	0.743
	Group B (With AI first)	Sensitivity	80.1% (73.7% - 85.9%)	78.2% (71.2% - 84.6%)	0.453
		Specificity	56.5% (48.6% - 64.5%)	58.7% (50.7% - 66.7%)	0.375
		Accuracy	69.1% (63.6% - 74.5%)	69.0% (63.6% - 74.1%)	0.146

Table 2 – Performance of Barrett non-experienced endoscopists with and without AI during the evaluation of video cases with non-dysplastic Barrett's or Barrett's related neoplasia.
Abbreviations – AI: Artificial intelligence-based clinical decision support system

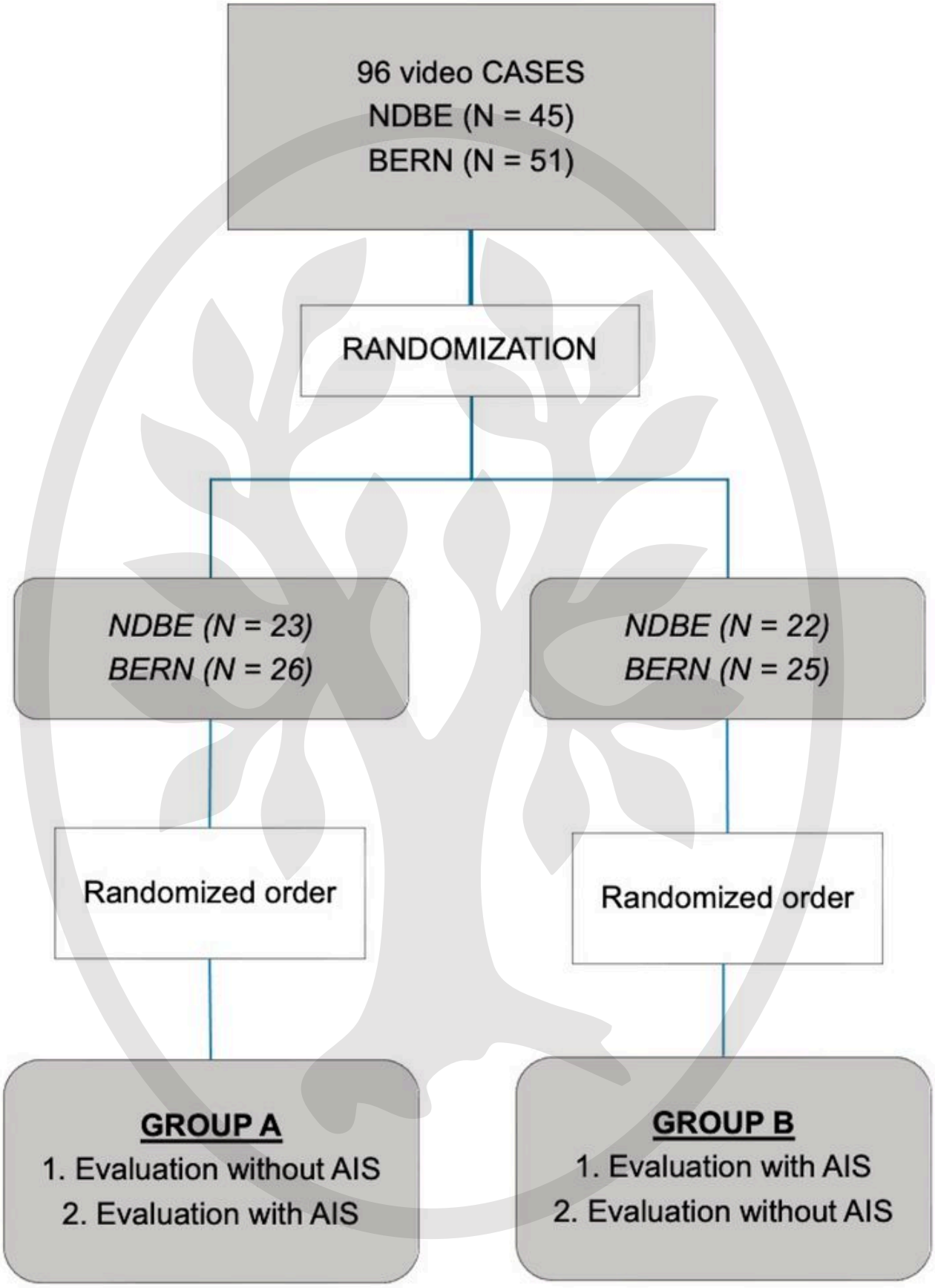
This article is protected by copyright. All rights reserved.
Accepted Manuscript

	STANDALONE PERFORMANCE OF AI (95% CI)
SENSITIVITY:	92.1 (88.2% - 95.6%)
SPECIFICITY:	68.9 (62.2% - 75.6%)
ACCURACY:	81.3 (77.3% - 85.2%)

Table 3 – Standalone performance of an AI system in the evaluation of Barrett's esophagus.
Abbreviations – AI: Artificial intelligence-based clinical decision support system

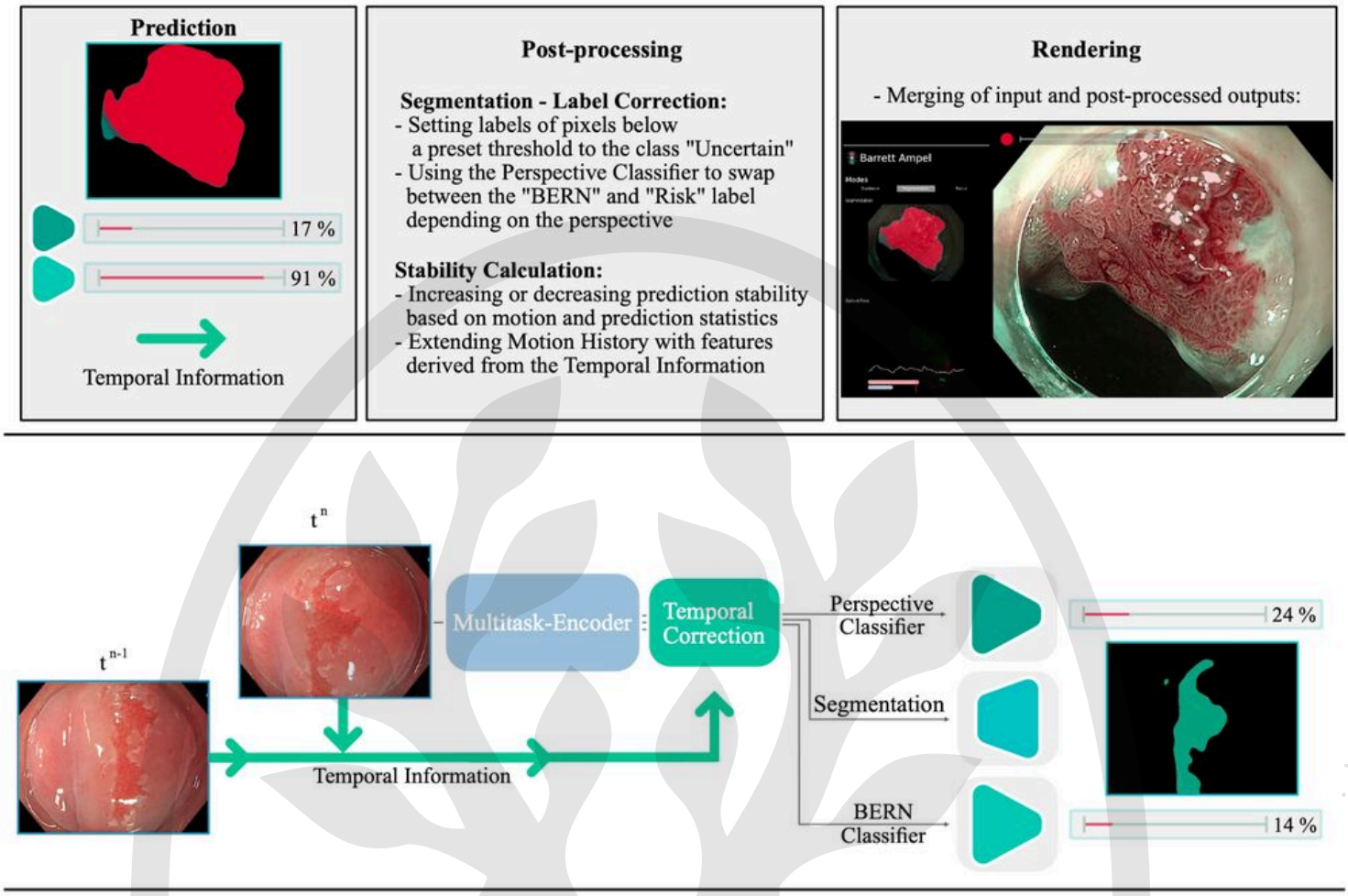
This article is protected by copyright. All rights reserved.

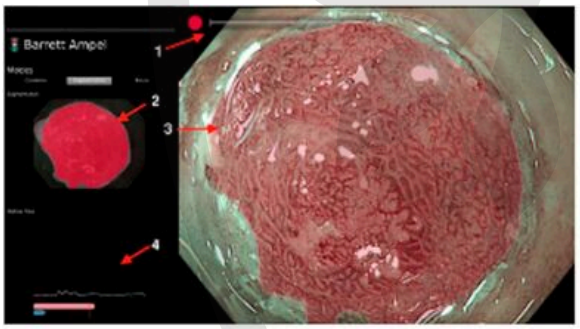
Accepted Manuscript



This article is protected by copyright. All rights reserved.

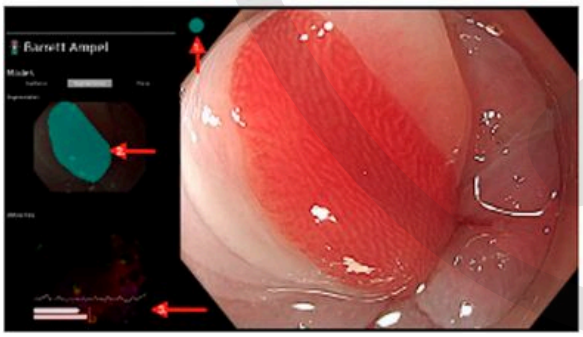
Accepted Manuscript





Suspicion of BE + BERN (confidence: high)

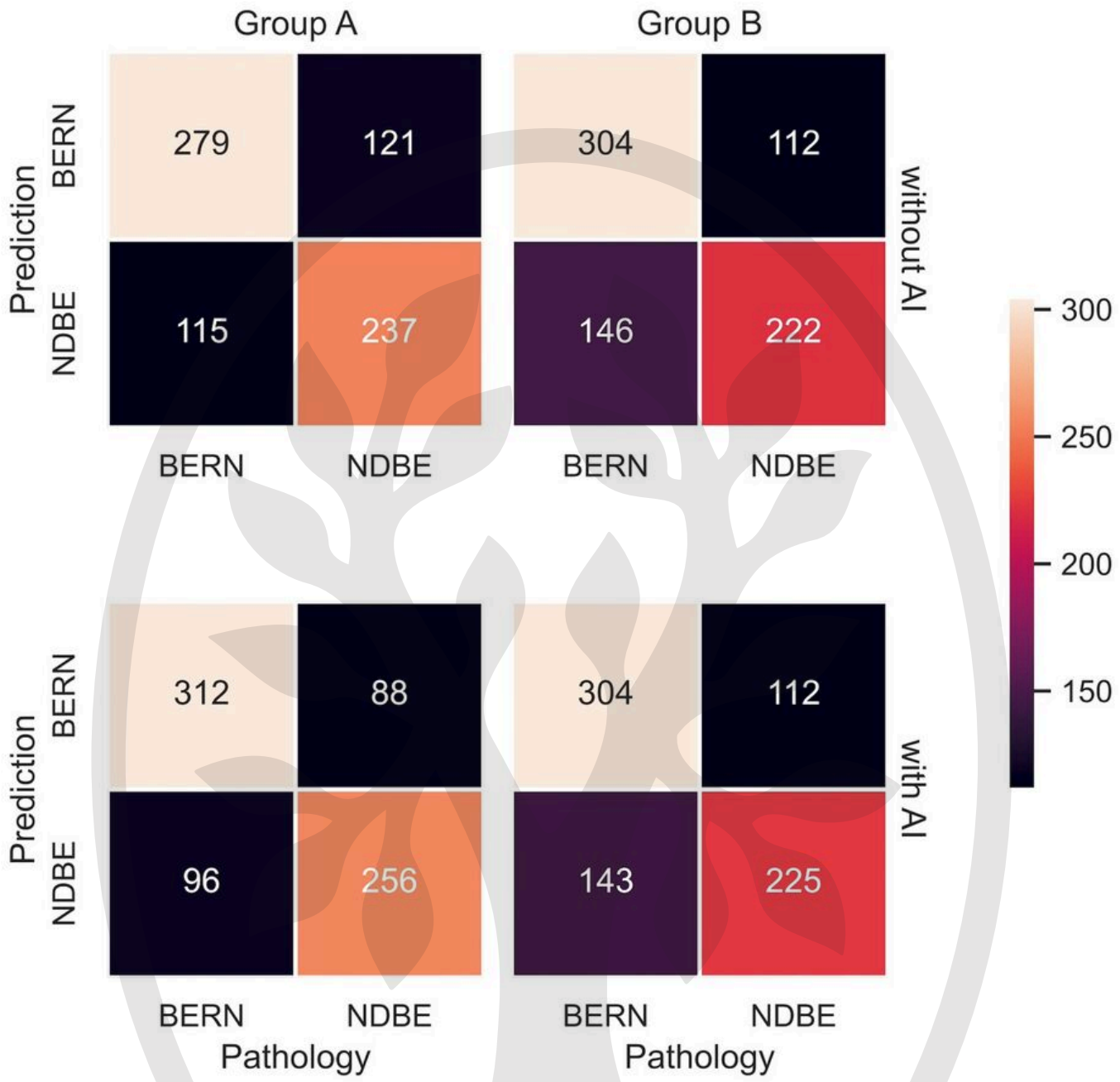
1. Global prediction/classification (red = BERN)
2. Estimate
3. Real-time segmentation
4. Optical Flow



BE without suspicion of BERN

1. Global prediction/classification (green = BE)
2. Suspicion of BE in the green delineated area
3. Optical Flow

This article is protected by copyright. All rights reserved.
Accepted Manuscript



This article is protected by copyright. All rights reserved.

Accepted Manuscript