A deep learning system for detection of early Barrett's neoplasia

Barrett's Oesophagus Imaging for Artificial Intelligence (BONS-AI) consortium; Fockens, K N; Jong, M R; Jukema, J B; Boers, T G W; Kusters, C H J; van der Putten, J A; Pouw, R E; Duits, L C; Montazeri, N S M

Published in:
Lancet digital health

DOI:
10.1016/S2589-7500(23)00199-1

Publication date:
2023

A deep learning system for detection of early Barrett's neoplasia: a model development and validation study

K N Fockens, M R Jong*, J B Jukema, T G W Boers, C H J Kusters, J A van der Putten, R E Pouw, L C Duits, N S M Montazeri, S N van Munster, B L A M Weusten, L Alvarez Herrero, M H M G Houben, W B Nagengast, J Westerhof, A Alkhalaf, R C Mallant-Hent, P Scholten, K Ragunath, S Seewald, P Elbe, F Baldaque-Silva, M Barret, J Ortiz Fernández-Sordo, G Moral Villarejo, O Pech, T Beyna, F van der Sommen, P H de With, A J de Groof, J J Bergman, on behalf of the Barrett's Oesophagus Imaging for Artificial Intelligence (BONS-AI) consortium†*

Summary

Background Computer-aided detection (CADe) systems could assist endoscopists in detecting early neoplasia in Barrett's oesophagus, which could be difficult to detect in endoscopic images. The aim of this study was to develop, test, and benchmark a CADe system for early neoplasia in Barrett's oesophagus.

Methods The CADe system was first pretrained with ImageNet followed by domain-specific pretraining with GastroNet. We trained the CADe system on a dataset of 14 046 images (2506 patients) of confirmed Barrett's oesophagus neoplasia and non-dysplastic Barrett's oesophagus from 15 centres. Neoplasia was delineated by 14 Barrett's oesophagus experts for all datasets. We tested the performance of the CADe system on two independent test sets. The all-comers test set comprised 327 (73 patients) non-dysplastic Barrett's oesophagus images, 82 (46 patients) neoplastic images, 180 (66 of the same patients) non-dysplastic Barrett's oesophagus videos, and 71 (45 of the same patients) neoplastic videos. The benchmarking test set comprised 100 (50 patients) neoplastic images, 300 (125 patients) non-dysplastic images, 47 (47 of the same patients) neoplastic videos, and 141 (82 of the same patients) non-dysplastic videos, and was enriched with subtle neoplasia cases. The benchmarking test set was evaluated by 112 endoscopists from six countries (first without CADe and, after 6 weeks, with CADe) and by 28 external international Barrett's oesophagus experts. The primary outcome was the sensitivity of Barrett's neoplasia detection by general endoscopists without CADe assistance versus with CADe assistance on the benchmarking test set. We compared sensitivity using a mixed-effects logistic regression model with conditional odds ratios (ORs; likelihood profile 95% CIs).

Findings Sensitivity for neoplasia detection among endoscopists increased from 74% to 88% with CADe assistance (OR 2·04; 95% CI 1·73–2·42; p<0·0001 for images and from 67% to 79% [2·35; 1·90–2·94; p<0·0001] for video) without compromising specificity (from 89% to 90% [1·07; 0·96–1·19; p=0·20] for images and from 96% to 94% [0·94; 0·79–1·11;] for video; p=0·46). In the all-comers test set, CADe detected neoplastic lesions in 95% (88–98) of images and 97% (90–99) of videos. In the benchmarking test set, the CADe system was superior to endoscopists in detecting neoplasia (90% *vs* 74% [OR 3·75; 95% CI 1·93–8·05; p=0·0002] for images and 91% *vs* 67% [11·68; 3·85–47·53; p<0·0001] for video) and non-inferior to Barrett's oesophagus experts (90% *vs* 87% [OR 1·74; 95% CI 0·83–3·65] for images and 91% *vs* 86% [2·94; 0·99–11·40] for video).

Interpretation CADe outperformed endoscopists in detecting Barrett's oesophagus neoplasia and, when used as an assistive tool, it improved their detection rate. CADe detected virtually all neoplasia in a test set of consecutive cases.

Funding Olympus.

Lancet Digit Health 2023; 5: e905–16

*Both authors contributed equally to this manuscript

†Members and collaborators of the BONS-AI consortium are listed in the appendix

Department of Gastroenterology and Hepatology, Amsterdam Gastroenterology, Endocrinology and Metabolism (K N Fockens MD, M R Jong MD, J B Jukema MD, R E Pouw MD PhD, L C Duits MD PhD, S N van Munster MD PhD, A J de Groof MD PhD, J J Bergman MD PhD), **and Biostatistics Unit, Department of Gastroenterology and Hepatology** (N S M Montazeri PhD), **Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands; Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands** (T G W Boers MEng, C H J Kusters MEng, J A van der Putten PhD, F van der Sommen PhD, P H de With PhD); **Department of Gastroenterology and Hepatology, UMC Utrecht, University of Utrecht, Utrecht, Netherlands** (B L A M Weusten MD PhD); **Department of Gastroenterology and Hepatology, St Antonius Hospital, Nieuwegein, Netherlands** (S N van Munster, B L A M Weusten, L Alvarez Herrero MD PhD); **Department of Gastroenterology and Hepatology, HagaZiekenhuis Den Haag, Den Haag, Netherlands** (M H M G Houben MD PhD); **Department of Gastroenterology and Hepatology, UMC Groningen, University of Groningen, Groningen, Netherlands**

Introduction

The endoscopic detection of early neoplasia in Barrett's oesophagus is challenging because early neoplastic lesions often have a subtle endoscopic appearance, with only minimal mucosal and vascular changes.[1,2] State-of-the-art endoscopes allow visualisation of nearly all such subtle changes, and, as a result, recognition of neoplasia by the endoscopist has become the rate-limiting factor in diagnosing early Barrett's neoplasia. Computer-aided detection (CADe) systems could assist endoscopists in this recognition. In recent years, multiple CADe systems for Barrett's neoplasia have been developed.[3–14] However, most of these studies rely on relatively small, retrospectively collected single-centre datasets, which restricts the generalisability of the results. Furthermore, most current CADe systems require substantial computational resources, limiting efficient integration in existing endoscopy systems.[15]

The BONS-AI consortium has recently developed an image-based CADe system for Barrett's neoplasia.[15] The BONS-AI consortium consists of 15 international centres with a tertiary referral function for the management of early Barrett's neoplasia. Endoscopic images of patients with neoplastic and non-dysplastic Barrett's oesophagus

(W B Nagengast MD PhD, J Westerhof MD PhD); **Department of Gastroenterology and Hepatology, Isala Hospital Zwolle, Zwolle, Netherlands** (A Alkhalaf MD PhD); **Department of Gastroenterology and Hepatology, Flevoziekenhuis Almere, Almere, Netherlands** (R C Mallant-Hent MD PhD); **Department of Gastroenterology and Hepatology, Onze Lieve Vrouwe Gasthuis, Amsterdam, Netherlands** (P Scholten MD); **Department of Gastroenterology and Hepatology, Royal Perth Hospital, Curtin University, Perth, WA, Australia** (K Ragunath MD PhD); **Department of Gastroenterology and Hepatology, Hirslanden Klinik, Zurich, Switzerland** (S Seewald MD PhD); **Department of Digestive Diseases, Karolinska University Hospital, Stockholm, Sweden** (P Elbe MD PhD, F Baldaque-Silva MD PhD); **Division of Surgery, Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Stockholm, Sweden** (P Elbe); **Center for Advanced Endoscopy Carlos Moreira da Silva, Gastroenterology Department, Pedro Hispano Hospital, Matosinhos, Portugal** (F Baldaque-Silva); **Department of Gastroenterology and Hepatology, Cochin Hospital Paris, Paris, France** (M Barret MD PhD); **Department of Gastroenterology and Hepatology, Nottingham University Hospitals NHS Trust, Nottingham, UK** (J Ortiz Fernández-Sordo MD, G Moral Villarejo MD); **Department of Gastroenterology and Hepatology, St John of God Hospital, Regensburg, Germany** (O Pech MD PhD); **Department of Gastroenterology and Hepatology, Evangalisches Krankenhaus Düsseldorf, Düsseldorf, Germany** (T Beyna MD PhD)

Research in context

Evidence before this study

We searched PubMed for publications in English from database inception to Jan 1, 2022, using the keywords "artificial intelligence" OR "deep learning" OR "computer-aided detection" AND "Barrett" OR "Barrett's neoplasia" OR "esophagus". In the last decade, several studies describe the development of preliminary computer-aided detection (CADe) systems for Barrett's neoplasia, generally reporting high diagnostic accuracy. However, most of these CADe systems were trained and tested on small, retrospective, or single-centre datasets and most studies have only reported stand-alone CADe performance.

Added value of this study

The CADe system described in this model development and validation study is developed using a large, heterogeneous dataset containing endoscopic images and video from 15 international endoscopy centres. This dataset contributes significantly to the robustness of our CADe system. This study is the largest AI study in the field of Barrett's neoplasia to date. The CADe system was tested on multiple external test sets. Benchmarking studies were performed to evaluate additive value of the CADe system when used by general endoscopists. The CADe system outperformed general endoscopists in detecting Barrett's oesophagus neoplasia and improved their detection rate, without compromising specificity. Furthermore, the system was specifically designed for direct implementation into current existing endoscopy platforms.

Implications of all the available evidence

Endoscopic detection of early Barrett's neoplasia is challenging because of its subtle appearance and relatively low incidence. CADe systems could potentially aid the endoscopist in neoplasia recognition. Several preliminary CADe systems for Barrett's neoplasia have been developed recently. However, for clinical implementation, CADe systems should be extensively trained and tested using large heterogeneous datasets and include benchmarking studies, followed by clinical studies. The robustness of CADe systems against data heterogeneity and real-world variability, as well as its additive value when used by general endoscopists, are essential for successful integration into routine clinical practice.

had been retrospectively collected in these centres. In parallel, a prospective standardised image and video acquisition protocol has been initiated across these centres. This large-scale heterogeneous data collection is essential for the development of a robust and reliable CADe system that is generalisable to daily clinical practice at other hospitals. We have recently described the infrastructure of the consortium and the performance of a preliminary image-based CADe system using only retrospectively collected images.[15]

In this study, we expanded the retrospective dataset and integrated prospectively collected data for training and testing of a CADe system for detecting Barrett's neoplasia using video. We aimed to compare the performance of the CADe system with the performance of Barrett's oesophagus experts and general, non-expert endoscopists to study whether the performance of general endoscopists improves when they are provided with CADe assistance and how their CADe-assisted performance compares to that of Barrett's oesophagus experts.

Methods

Study design

This model development and validation study was conducted by the Barrett's Oesophagus Imaging for Artificial Intelligence (BONS-AI) consortium. The consortium is led by the Department of Gastroenterology and Hepatology of the Amsterdam University Medical Center, Amsterdam, Netherlands, and the Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands. We have recently described the composition of the BONS-AI consortium, and its infrastructure regarding data acquisition, endoscopic protocol, definitions of neoplastic and non-dysplastic images and video, creation of an expert-based delineation gold standard, and basic infrastructure of the CADe system.[15] We developed a CADe system for the primary detection of Barrett's neoplasia on endoscopic images and video. The system classifies images as either neoplastic or non-dysplastic, followed by localisation of neoplasia (if present) with a green bounding box around the lesion.[15]

The Medical Research Involving Human Subjects Act did not apply to this study. Therefore, official approval was waived by the medical ethics review committee of all participating centres. Although the need for a formal informed consent was waived by the institutional review boards, all patients were informed that de-identified images and video were being recorded for studies during prospective acquisition.

Pretraining the CADe system

We used the ImageNet dataset for generic pretraining of the CADe system.[16] ImageNet is a publicly available dataset of 1 200 000 general images with 1000 different categories, including buildings, animals, and vehicles. In this pretraining phase, the deep learning system learns basic features of images (such as edges and shapes). This knowledge is then transferred to the next phase of the training process. Such pretraining eliminates the need to learn basic features from images within the final application, which are often scarce for medical applications. After generic pretraining on ImageNet,

we performed domain-specific pretraining using GastroNet data. GastroNet is a dataset consisting of 5 084 494 unlabelled endoscopic images, developed by our consortium. General endoscopic images were retrieved from the endoscopic databases of Dutch hospitals (Amsterdam University Medical Center, University Medical Center Utrecht, University Medical Center Groningen, Catharina Ziekenhuis Eindhoven, Spaarne Gasthuis Haarlem, Medisch Spectrum Twente, Flevoziekenhuis Almere, and Isala Ziekenhuis Zwolle). Endoscopic images were recorded between Jan 1, 2012, and Dec 31, 2021. In previous work, we demonstrated that domain-specific pretraining improves the performance of deep learning systems in gastrointestinal endoscopy.[7]

Data for Barrett's neoplasia-specific training and validation

After pretraining, the CADe system underwent refinement training with retrospectively and prospectively collected endoscopic images of patients with Barrett's oesophagus.

Retrospectively collected images were recorded between Jan 1, 2012, and Dec 31, 2021, using Olympus H180, HQ190, and HQ290 endoscopes and CV180, CV190, and CV290 processors (Olympus Europa, Hamburg, Germany). Eligible images were identified using hospital-specific patient lists containing data on patients under Barrett's oesophagus surveillance or patients treated for early Barrett's neoplasia. Images were extracted from the endoscopy databases of 15 collaborative partners using anonymisation software, specifically designed for this project.[15]

In addition to these retrospectively collected images, endoscopic images and video were prospectively collected in all participating centres following a standardised acquisition protocol. All images and video were recorded without a specific focus on a neoplastic lesion, if present. At 2 cm intervals throughout the Barrett's segment and in the retroflexed position, a 10 s overview video was recorded, followed by two endoscopic images. Prospective endoscopic imagery was obtained using Olympus HQ190 and EZ1500 gastroscopes and Olympus CV190 and X1 processors. Images and video were recorded using MediCap USB300 (MediCapture, Plymouth Meeting, PA, USA) for CV190 processors and HVO-4000MT (Sony Corporation, Tokyo, Japan) for X1 processors. A detailed description of the prospective acquisition protocol has been published elsewhere.[15] Neoplastic and non-dysplastic images were drawn from the same patient distribution (eg, patients visiting hospital) and had similar properties (eg, distal attachment cap usage, endoscope type).

The final training set consisted of 7595 non-dysplastic Barrett's oesophagus images derived from 1095 patients (5228 retrospective and 2367 prospective images) and 6251 neoplastic images derived from 1296 patients (4915 retrospective and 1336 prospective images; table 1).

Optimising thresholds, parameters, and hyperparameters

To optimise the performance of the CADe system, the system was validated on 200 images of Barrett's oesophagus that were prospectively collected. Based on this dataset, hyperparameters of the CADe system were optimised and an optimal threshold to alert for neoplasia was selected. It is important to note that this threshold value was not calibrated and, therefore, not related to disease incidence or to a specific chance of lesion prevalence in the image. This dataset consisted of 100 neoplastic images (derived from 58 patients) and 100 non-dysplastic Barrett's oesophagus images (derived from 36 patients). To specifically improve the system's capability of detecting subtle neoplasia, we enriched this

Correspondence to:
Dr J J Bergman, Department of Gastroenterology and Hepatology, Amsterdam Gastroenterology, Endocrinology and Metabolism, Amsterdam UMC, University of Amsterdam, Amsterdam 1007, Netherlands
j.j.bergman@amsterdamumc.nl

See Online for appendix

	Number of images; patients	Neoplastic images; patients	NDBE images; patients	Acquisition	Type of labelling
ImageNet	1 200 000; NA	NA	NA	NA	NA
GastroNet	5 084 494; unknown	NA	NA	Retrospective acquisition	Subset: hand-labelled by two experts
Training set	13 846; 2391	6251; 1296	7595; 1095	Retrospective and prospective acquisition	Hand-labelled by three experts, correlating pathology, delineated by two or more experts
Image validation set	200; 94	100; 58	100; 36	Prospective acquisition	Hand-labelled by three experts, correlating pathology, delineated by two or more experts
Video validation set	180; 91	77; 58	103; 33	Retrospective and prospective acquisition	NA
All-comers image test set	409; 119	82; 46	327; 73	Prospective acquisition	Hand-labelled by three experts, correlating pathology, delineated by two or more experts
All-comers video test set	251; 111	71; 45	180; 66	Prospective acquisition	Hand-labelled by three experts, correlating pathology, delineated by two or more experts
Benchmarking image test set	400; 175	100; 50	300; 125	Prospective acquisition	Hand-labelled by three experts, correlating pathology
Benchmarking video test set	188; 129	47; 47	141; 82	Prospective acquisition	Hand-labelled by three experts, correlating pathology

CADe=computer-aided detection. NA=not applicable. NDBE=non-dysplastic Barrett's oesophagus.

Table 1: **Overview of used datasets for the development of the CADe system**

dataset with cases of small or flat lesions. From the same patient group, an additional 77 neoplastic videos and 103 non-dysplastic Barrett's oesophagus videos were used for video-specific optimisation.

Prospective datasets to test CADe performance

We evaluated CADe performance on two independent test sets. We developed an all-comers test set to evaluate the performance of the CADe system on cases representing daily clinical practice. This test set consisted of images and video of consecutive cases prospectively recorded at the BONS-AI consortium centres between Jan 1, 2022, and March 1, 2022. The all-comers image test set consisted of 327 non-dysplastic Barrett's oesophagus images from 73 patients and 82 neoplastic images from 46 patients. The corresponding video test set consisted of 180 non-dysplastic Barrett's oesophagus videos from 66 patients and 71 neoplastic videos from 45 patients (table 1).

We developed a benchmarking test set to evaluate the performance of the CADe system on more challenging cases. We artificially enriched this test set with subtle neoplasia, since these cases are the most challenging for endoscopists in daily practice. Lesion subtlety was independently determined by two experts (JJB and AJdG). The image benchmarking test set comprised 400 Barrett's oesophagus images: 100 neoplastic images from 50 patients and 300 non-dysplastic images from 125 patients. The corresponding video test set originated from the same patient group and comprised 188 Barrett's oesophagus videos: 47 neoplastic videos from 47 patients and 141 non-dysplastic videos from 82 patients (table 1). Further information on the test sets, in terms of pathology and macroscopic appearance, is described in the appendix (p 8). Both test sets were stored at a separate site for a single performance test. There was no overlap in patients or images between the test sets and the training set or validation set.

Ground truth development

To provide a ground truth classification and segmentation for training the CADe system, 4793 neoplastic images were delineated by 14 expert endoscopists (REP, BLAMW, MHMGH, WBN, JW, LAH, AA, KR, MB, JOF-S, OP, TB, SS, and JJB) using proprietary software (Meducati, Göteborg, Sweden). Experts were defined as having a scientific track record (ie, authored on >10 peer-reviewed studies) and clinical track record (ie, working for >5 years in a tertiary referral centre) for diagnosis and treatment of early Barrett's neoplasia. Each image was delineated by at least two experts. Based on overlapping areas between the delineations of these two experts, the ground truth for neoplasia was determined. Details on the delineation process of neoplastic images and videos have been published elsewhere[15] and are described in the appendix (p 3).

All images and video were reviewed by three research fellows (KNF, JBJ, and MRJ) to ensure the standard of photographic quality, the absence of any visible abnormalities in the non-dysplastic Barrett's oesophagus group, and the presence of a visible neoplastic lesion in the neoplastic group. Non-dysplastic Barrett's oesophagus images were only included if all biopsies obtained at the corresponding endoscopy showed the absence of dysplasia of any grade and if there had been no previous endoscopic treatment of the Barrett's segment. The images in the neoplastic group all contained a visible lesion with high-grade dysplasia or adenocarcinoma in the corresponding endoscopic resection specimen since they underwent endoscopic resection in the same or follow-up session. All neoplastic images were obtained from patients who were treatment naive for Barrett's neoplasia.

Benchmark performance by expert endoscopists

To provide a reference for the performance of the CADe system, 28 Barrett's oesophagus experts, originating from six countries and all with an international reputation in the endoscopic management of Barrett's oesophagus, were invited to evaluate the benchmarking test set. None of the Barrett's oesophagus experts were affiliated with the BONS-AI consortium and, therefore, all cases were completely new to them. A previously designed web-based module (Meducati, Göteborg, Sweden)[15,17,18] was adjusted and used for this specific study.

14 of 28 Barrett's oesophagus experts evaluated the benchmarking image test set. Due to the extent of the test set, it was divided into two parts, each containing 50 neoplastic images originating from 50 patients and 150 randomly selected non-dysplastic images. Each Barrett's oesophagus expert evaluated one of the two subsets. The other 14 Barrett's oesophagus experts evaluated the benchmarking video test set. Each expert evaluated the complete test set of 188 videos, which included 47 neoplastic and 141 non-dysplastic Barrett's oesophagus cases. The order of images and videos in all sets was randomised per expert.

The Barrett's oesophagus experts scored every case for the presence of neoplasia. If a neoplastic lesion was detected, the expert was asked to place a targeted biopsy mark on the most abnormal area of the lesion. This place was supposed to represent the location for targeted biopsy during real-time endoscopic examination. The assessment of videos is described in the appendix (pp 3–4). All images and videos were evaluated without any information on patient status.

Two-phase benchmark performance by general endoscopists

The CADe system was developed to assist general endoscopists without specific Barrett's oesophagus expertise. To provide an appropriate reference for the performance of the CADe system, and to evaluate the additive value of the CADe system for general

endoscopists, we invited a large and heterogeneous group of endoscopists to evaluate either the benchmarking image test set or the benchmarking video test set. The endoscopists originated from six different countries and had varying levels of endoscopic experience, divided accordingly into three groups: less than 3 years, 3–10 years, and more than 10 years of endoscopic experience.

This benchmarking process consisted of two phases. The first phase was identical to the module of the expert group described. To mitigate a learning effect or a biased assessment, the second phase was initiated after a 6-week washout period. It comprised the assessment of the same images or videos as in phase 1, in a different, randomised order with CADe assistance. The endoscopists evaluated the images and video for the presence of neoplasia; however, during the second phase, the CADe prediction was projected as a green bounding box over areas of suspected neoplasia (figure 1). Subsequently, the placement of targeted biopsies was identical to phase one. No feedback was given to the endoscopists between assessment phases.

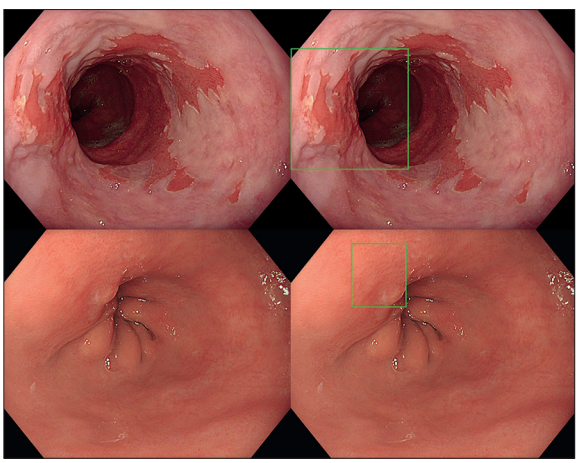

Figure 1: **Visualisation prediction by the computer aided detection system using a green bounding box**

Architecture of the CADe system

The CADe system was constructed using an EfficientNet-Lite1 encoder[19] to extract the relevant image features and a MobileNetV2 DeepLabV3+ decoder[20] to generate an output segmentation. Both architectures were further optimised for efficient execution on many existing embedded hardware platforms that are typically used in endoscopy systems.

Following the two-step pretraining sequence with ImageNet and GastroNet, the encoder and decoder (classification and segmentation) branch of the system were trained simultaneously using the Barrett's oesophagus training dataset to classify and localise Barrett's neoplasia. During validation, the best performance was achieved using only the segmentation branch of the decoder for both classification and segmentation. To ensure the best and most consistent outcomes, we used the segmentation branch for both classification and segmentation. Roughly half of all neoplastic images had expert delineations. Neoplastic images without expert delineation could still be efficiently leveraged to improve the training of the encoder using a small classification head on the bridge of the architecture. Additional technical details are described in the appendix (pp 5–6).

Performance metrics of CADe and endoscopists

Classification was considered correct when the endoscopist and the CADe system correctly classified an image or video as neoplastic or non-dysplastic Barrett's oesophagus using the histological label of the image or video as the gold standard. For the video stand-alone performance on classification, the CADe assessment was considered neoplastic whenever the system classified 75% or more of the frames as neoplastic during a time interval of 1·5 s in a post-hoc calculation. This time interval was determined based on our internal validation set. We rationalised these cutoff values because endoscopic videos consist of thousands of individual frames. If the detection of a single frame in a video would count as a positive detection, the video-based outcomes would always result in 100% sensitivity and 0% specificity. Furthermore, brief detections of a single frame are hardly visible to the human eye. Therefore, clinically relevant detections of Barrett's neoplasia should comprise sequential positive frames over a defined amount of time.

We displayed classification performance in terms of sensitivity and specificity for each of the test sets used. Given the clear differences in clinical relevance of sensitivity and specificity, we did not report accuracy.

Localisation scores were calculated for correctly classified neoplastic imagery and were defined as the proportions of neoplastic images or videos in which the lesion was correctly localised by either the endoscopist or the CADe system. For endoscopists, localisation was considered correct when the biopsy marker was correctly placed within the ground truth area for both images and videos. For the CADe system, localisation on images was considered correct when the bounding box overlapped the ground truth area. No localisation score of the CADe system was calculated for videos as there was no available ground truth for every frame.

Outcomes

The primary outcome was classification performance of general endoscopists in phase one versus performance in phase two (when assisted by the CADe system [for images and for videos]). We report this outcome with sensitivity. Secondary outcomes were the classification performance of general endoscopists in phase one versus performance in phase two when assisted by the CADe system (for images and for videos), reported with specificity; stand-alone classification performance of the CADe system on

	Classification		Localisation	
	Sensitivity	Specificity	Performance	Method
All-comers image test set				
CADe	95% (88–98)	70% (65–75)	100% (95–100)	Bounding box
All-comers video test set				
CADe	97% (90–99)	85% (79–89)	NA	NA
Benchmarking image test set				
CADe	90% (83–94)	80% (75–84)	100% (96–100)	Bounding box
General endoscopists	74% (66–84; 73–76)	89% (83–91; 85–86)	92% (89–92)	Biopsy mark
General endoscopists and CADe	88% (74–92; 82–85)	90% (83–93; 85–87)	92% (91–93)	Biopsy mark
Expert endoscopists	87% (76–93; 82–88)	86% (77–95; 83–86)	94% (91–95)	Biopsy mark
Benchmarking video test set				
CADe	91% (80–97)	82% (74–87)	NA	NA
General endoscopists	67% (58–79; 65–70)	96% (91–99; 92–94)	100% (92–95)	Biopsy mark
General endoscopists and CADe	79% (67–92; 75–79)	94% (91–97; 92–94)	96% (93–96)	Biopsy mark
Expert endoscopists	86% (79–94; 80–86)	90% (87–96; 87–90)	96% (93–97%)	Biopsy mark

Data are in median (IQR; 95% CI), unless stated. CADe scores are presented as specific values (95% CI). CADe=computer-aided detection. NA=not applicable.

Table 2: **Performance of the CADe system and endoscopists on all test sets**

	Sensitivity		Specificity	
	OR (95% CI)	p value	OR (95% CI)	p value
Benchmarking image test set				
CADe *vs* general endoscopists	3·75 (1·93–8·05)	0·0002	0·64 (0·47–0·89)	0·0047
General endoscopists with CADe *vs* general endoscopists	2·04 (1·73–2·42)	<0·0001	1·07 (0·96–1·19)	0·20
CADe *vs* expert endoscopists	1·74 (0·83–3·65)	0·14	0·70 (0·50–0·98)	0·032
General endoscopists with CADe *vs* expert endoscopists	0·85 (0·38–1·91)	0·69	1·13 (0·52–2·42)	0·75
Benchmarking video test set				
CADe *vs* general endoscopists	11·68 (3·85–47·53)	<0·0001	0·30 (0·19–0·49)	<0·0001
General endoscopists with CADe *vs* general endoscopists	2·35 (1·90–2·94)	<0·0001	0·94 (0·79–1·11)	0·46
CADe *vs* expert endoscopists	2·94 (0·99–11·40)	0·078	0·54 (0·34–0·89)	0·012
General endoscopists with CADe *vs* expert endoscopists	0·42 (0·16–1·03)	0·055	1·87 (1·58–2·21)	0·050

CADe=computer-aided detection. OR=odds ratio.

Table 3: **Results of mixed-model analysis**

the all-comers test set (for images and for videos), reported with sensitivity and specificity; stand-alone classification performance of the CADe system, the Barrett's oesophagus experts, and the general endoscopists on the benchmarking test set (for images and for videos), reported with sensitivity and specificity; and localisation performance of the CADe system, Barrett's oesophagus experts, and general endoscopists on the benchmarking image test set and, for Barrett's oesophagus experts and general endoscopists, on the benchmarking video test set.

For the benchmarking test set, we aimed to assess four clinically relevant comparisons: (1) stand-alone performance of CADe versus the performance of general endoscopists in phase one; (2) difference in the performance of general endoscopists between phase one (no CADe assistance) and phase two (with CADe assistance); (3) stand-alone performance of CADe versus the performance of the Barrett's oesophagus experts; and (4) performance of general endoscopists with CADe assistance versus the performance of Barrett's oesophagus experts. For comparisons (3) and (4), we aimed to evaluate non-inferiority for neoplasia detection. For the all-comers test set, we considered stand-alone CADe performance as an exploratory result and, therefore, we only reported sensitivity and specificity.

Statistical analysis

For descriptive statistics, we show the median (IQR) for continuous variables and frequencies with percentages for categorical variables. CIs were obtained using the Wilson method. Performance metrics were calculated using Python (3.8.10). Statistical analyses were done in R Studio (4.2.1).

We compared sensitivity using a mixed-effects logistic regression model (lme4 package) in a subset of neoplastic images or videos only, since sensitivity is defined as the number of true positives (neoplastic images or videos that were detected by the specific group of interest), divided by all positives (ie, all neoplastic images or videos in the dataset). This comparison resulted in a conditional odds ratio (OR) for sensitivity, along with likelihood profile 95% CIs. For specificity, we used a similar model in the subset of non-dysplastic images or videos only, with specificity defined as the number of true negative results (non-dysplastic images or videos that were defined as flat Barrett's oesophagus without neoplastic lesion by the specific group of interest) divided by all negative results (ie, all non-dysplastic images or videos in the dataset). For clinical comparisons (3) and (4), we aimed to evaluate non-inferiority for neoplasia detection. The non-inferiority margin was set at –5% sensitivity in comparison to experts, with an expected expert sensitivity of 90% based on internal data (unpublished). For more information on the mixed-effect logistic regressions models and non-inferiority analysis, see the appendix (pp 4–5). This study was registered at the Dutch Trial Register, NL8411.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Results

We evaluated the performance of the CADe system during training using the validation dataset. Based on this dataset, we optimised the parameters and hyper-parameters and selected the threshold for neoplasia. Striving for high sensitivity, we selected a threshold of 0·35, indicating that a CADe prediction score of 0·35 or

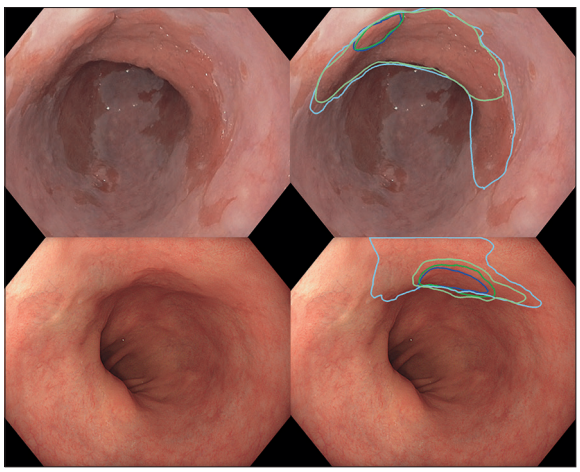

Figure 2: **Neoplastic lesions missed by the computer-aided detection system with corresponding expert ground truth**

more classified the image or video frame as neoplastic, whereas a prediction score of less than 0·35 classified the image or video frame as non-dysplastic (appendix p 10).

In the all-comers image test set, the CADe system correctly classified 78 of 82 neoplastic images (95% sensitivity) and 229 of 327 non-dysplastic images (70% specificity). In the all-comers video test set, the CADe system correctly classified 69 of 71 neoplastic videos (97% sensitivity) and 153 of 180 non-dysplastic videos (85% specificity). There was no overlap of missed cases between the all-comers image test set and the all-comers video test set. Therefore all lesions were detected at least in one modality.

In the benchmarking image test set, the CADe system correctly classified 90 of 100 neoplastic images (90% sensitivity, 95% CI 83–94) and 240 of 300 non-dysplastic images (80% specificity, 75–84). In the benchmarking video test set, the CADe system correctly classified 43 of 47 neoplastic videos (sensitivity 91%, 80–97) and 115 of 141 non-dysplastic videos (specificity 82%, 74–87; tables 2, 3). Similar to the all-comers test sets, there was no overlap of missed cases (figure 2) between the benchmarking image test set and the benchmarking video test set. Therefore, all lesions were detected in at least one modality. In the benchmarking study, 49 endoscopists (image test set) and 63 endoscopists (video test set) completed both phases. The endoscopists originated from six countries. Details on the country of origin and their endoscopic expertise are shown in table 4.

For the benchmarking image test set, the median sensitivity of general endoscopists in phase one was 74% (IQR 66–84; figure 3A) and specificity was 89% (83–91). Sensitivity of the CADe system was superior to general endoscopists (90% *vs* 74%; OR 3·75 [95% CI 1·93–8·05]; p=0·0002). Specificity of CADe was inferior (80% *vs* 89%; 0·64 [0·47–0·89]; p=0·0047). Barrett's oesophagus experts had a median sensitivity of 87% (IQR 76–93) and a specificity of 86% (77–95; figure 3B). In terms of sensitivity, CADe was non-inferior to expert endoscopists (90% *vs* 87%; OR 1·74 [95% CI 0·83–3·65]). The specificity of CADe was lower than expert endoscopists (80% *vs* 86%; 0·70 [0·50–0·98]; p=0·032).

	Sensitivity	Specificity	Localisation
Benchmarking image test set evaluated by general endoscopist			
All general endoscopists	74% (73–76%)	89% (85–86)	92% (89–92)
Experience			
<3 years (14)	68% (67–74)	85% (83–86)	90% (85–91)
3–10 years (20)	74% (73–78)	90% (86–88)	92% (89–93)
>10 years (15)	76% (75–81)	86% (82–85)	92% (89–93)
Country of origin			
Netherlands (26)	75% (70–75)	89% (87–89)	85% (90–93)
Germany (20)	75% (76–81)	85% (81–84)	83% (88–92)
Australia (3)	70% (67–81)	77% (78–85)	76% (81–93)
Benchmarking image test set evaluated by general endoscopist with CADe assistance			
All	88% (82–85)	90% (85–87)	92% (91–93)
Experience			
<3 years (14)	79% (78–84)	91% (88–90)	93% (90–94)
3–10 years (20)	89% (82–86)	91% (86–88)	92% (90–93)
>10 years (15)	92% (81–86)	86% (81–84)	92% (90–94)
Country of origin			
Netherlands (26)	82% (79–84)	91% (87–89)	88% (92–95)
Germany (20)	91% (82–86)	88% (84–87)	87% (89–93)
Australia (3)	94% (83–93)	75% (69–77)	80% (81–92)
Benchmarking video test set evaluated by general endoscopist			
All	67% (65–70)	96% (92–94%)	100% (92–95)
Experience			
<3 years (28)	67% (64–71)	94% (91–93)	100% (90–95)
3–10 years (23)	67% (73–78)	96% (86–88)	100% (95–98)
>10 years (12)	73% (75–81)	95% (82–85)	91% (88–95)
Country of origin			
Netherlands (27)	67% (63–70)	97% (95–96)	100% (93–97)
Belgium (16)	79% (69–77)	94% (91–94)	95% (91–97)
USA (12)	65% (56–68)	95% (88–92)	97% (89–96)
Other (8)	67% (58–71)	96% (88–93)	94% (82–93)
Benchmarking video test set evaluated by general endoscopist with CADe assistance			
All	79% (75–79)	94% (92–94)	96% (92–95)
Experience			
<3 years (28)	83% (76–82)	94% (90–93)	95% (91–96)
3–10 years (23)	75% (72–79)	94% (94–96)	100% (93–97)
>10 years (12)	77% (70–80)	96% (90–94)	93% (91–97)
Country of origin			
Netherlands (27)	79% (72–78)	96% (92–94)	100% (94–98)
Belgium (16)	83% (75–83)	94% (93–95)	93% (90–96)
USA (12)	69% (68–78)	95% (87–91)	98% (91–97)
Other (8)	90% (79–89)	94% (92–96)	95% (86–95)

Data are in median (95% CI). CADe=computer-aided detection.

Table 4: **Results of general endoscopists by subgroup**

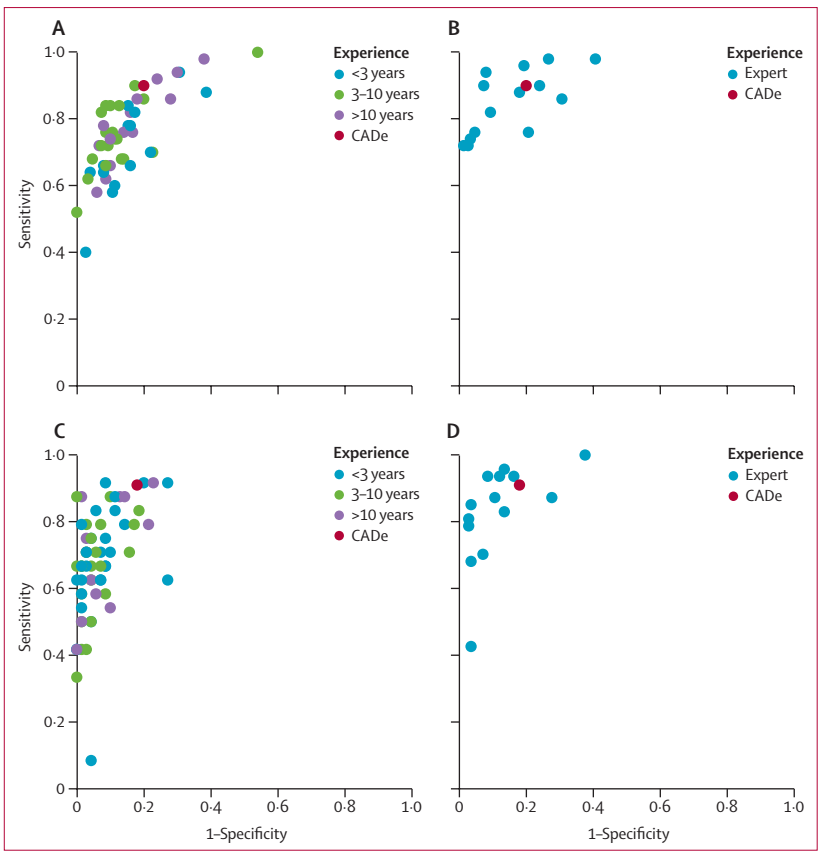

Figure 3: **Individual results of benchmarking test sets**
(A) General endoscopists on images. (B) International Barrett's oesophagus experts on images. (C) General endoscopists on video. (D) International Barrett's oesophagus experts on video. CADe=computer-aided detection.

For the benchmarking video test set, the median sensitivity of general endoscopists in phase one was 67% (IQR 58–79; figure 3C). Sensitivity of the CADe system was superior to the general endoscopists (91% *vs* 67%; OR 11·68 [95% CI 3·85–47·53]; p<0·0001). Specificity of CADe was inferior (82% *vs* 96%; 0·30 [0·19–0·49]; p<0·0001). The 14 Barrett's oesophagus experts had a median sensitivity of 86% (IQR 79–94%) with a specificity of 90% (IQR 87–96%; figure 3D). In terms of sensitivity, performance of the CADe system was non-inferior to expert endoscopists (91% *vs* 86%; OR 2·94 [95% CI 0·99–11·40]). As for specificity, CADe was inferior (82% *vs* 90%; 0·54 [0·34–0·89]; p=0·012).

On the benchmarking image test set, the sensitivity of general endoscopists increased significantly with CADe assistance, from 74% in phase one to 88% in phase two (OR 2·04 [95% CI 1·73–2·42]; p<0·001). For the specificity of general endoscopists, which was 89% in phase one and 90% in phase two, no significant difference was found between the two phases (OR 1·07 [95% CI 0·96–1·19]; p=0·20; figure 4). For the sensitivity of the general endoscopists with CADe assistance, non-inferiority to the Barrett's oesophagus experts could not be proven (88% *vs* 87%; OR 0·85 [95% CI 0·38–1·91]). As for specificity, no significant difference was found between the general endoscopists with CADe assistance and the Barrett's oesophagus experts (90% *vs* 86%; OR 1·13 [0·52–2·42]; p=0·75).

In the benchmarking video test set, the sensitivity of general endoscopists increased significantly with CADe assistance, from 67% in phase one to 79% in phase two (OR 2·35 [95% CI 1·90–2·94]; p<0·0001). For specificity, no significant difference was found between general endoscopists with or without CADe assistance, with 96% in phase one and 94% in phase two (0·94 [0·79–1·11]; p=0·46; figure 4). For sensitivity of general endoscopists with CADe on the benchmarking video test set, non-inferiority to Barrett's oesophagus experts again could not be shown (79% *vs* 86%; OR 0·42 [0·16–1·03]). As for specificity, general endoscopists with CADe assistance were superior to Barrett's oesophagus experts (94% *vs* 90%; OR 1·87 [1·58–2·21]; p=0·050).

In the benchmarking image test set, the CADe system correctly localised 100% of the neoplastic lesions in images that were correctly classified. General endoscopists, both with and without CADe assistance, had a median 92% localisation score, whereas expert endoscopists correctly localised 94% of all correctly classified lesions. In the video benchmarking test set, the general endoscopists had a median localisation score of 100% without CADe assistance and 96% with CADe assistance, whereas expert endoscopists correctly localised 96% of all correctly classified lesions.

Discussion

In this study, we developed, tested, and benchmarked a CADe system for Barrett's neoplasia using white light endoscopy on both images and videos. The main problem in endoscopic recognition of early Barrett's neoplasia is that neoplasia is often subtle and, therefore, easily missed by general endoscopists.

Since the aim of our CADe system was to increase the detection rate of neoplastic lesions, our primary endpoint was sensitivity. If the CADe increases sensitivity, specificity becomes important. The clinical consequences of missing a neoplastic lesion (low sensitivity) could be substantial as surveillance intervals can be up to 5 years and missing a neoplastic lesion in a 3 cm Barrett's segment could cause a delay in diagnosis with prognostic consequences. In contrast, overcalling a normal area as neoplastic (low specificity) will—at worst—lead to an additional targeted biopsy being obtained. The number of false-positive results should, logically, be within an acceptable range; however, we could accept some loss in specificity since the clinical consequence of false-positive results is limited and the majority false-positive detections by the CADe system are easily dismissed by the endoscopists, as we found in a previous study.[15] Furthermore, any positive detection could be further interrogated by a secondary computer-aided diagnosis

system for detailed characterisation, dismissing additional false-positive detections.

We found that CADe assistance significantly improved the image and video-based detection of early Barrett's neoplasia by general endoscopists. Without CADe assistance, the endoscopists missed a substantial number of high-grade dysplasia or adenocarcinoma lesions—eg, 53% of endoscopists missed more than 25% of lesions and a subgroup of 17% had a miss rate of more than 40%. These miss rates are similar to previous studies.[1,2,7]

The CADe system was tested with 112 general endoscopists from six countries, who had varying levels of endoscopic expertise. All subgroups based on country of origin and the years of endoscopic experience displayed a similar performance.

In this study, two test sets were used to study CADe performance. The all-comers test set consisted of consecutive cases presented in a timeframe of 2 months at the 15 participating centres of the BONS-AI consortium. The neoplastic lesions in this test set ranged from subtle to more obvious and CADe detected virtually all lesions. The sensitivity was 95% for images and 97% for videos and, in combination (ie, image and [or] video), all neoplastic lesions were detected. In the benchmarking test set the case mix was artificially enriched for subtle neoplasia as subtle neoplasia cases have a higher probability of being missed in daily practice by general endoscopists. In the benchmarking test set, the CADe system had a lower sensitivity (90% for images and 91% for videos) compared with in the all-comers test set. The missed neoplastic lesions were all subtle in appearance (figure 2); however, all were detected on either the still image or the corresponding video. These results resemble the intended application of this CADe system, in which the endoscopist starts with a general inspection of the Barrett's segment using real-time video-based CADe assistance. Subsequently, still images of each level might contribute to an additional CADe-assisted inspection of the Barrett's segment.

Although our CADe system significantly outperformed general endoscopists in terms of neoplasia detection, the system's specificity was lower (80% *vs* 89% for images and 82% *vs* 96% for videos). In earlier studies, we found that false-positive CADe detections can be separated into three subgroups: obvious false-positive detections caused by bubbles or light-reflections, subtle abnormalities that deserve targeted reinspection, and flat-type mucosa without apparent abnormalities. We speculated that most false-positive CADe detections could easily be discarded by endoscopists. In this study, we observed this synergistic effect between endoscopists and CADe assistance and, with CADe assistance, the general endoscopists increased their detection of early neoplasia while maintaining their own level of specificity.

Although endoscopists improved their neoplasia detection with CADe assistance, many of them did not reach the level of neoplasia detection of the CADe system

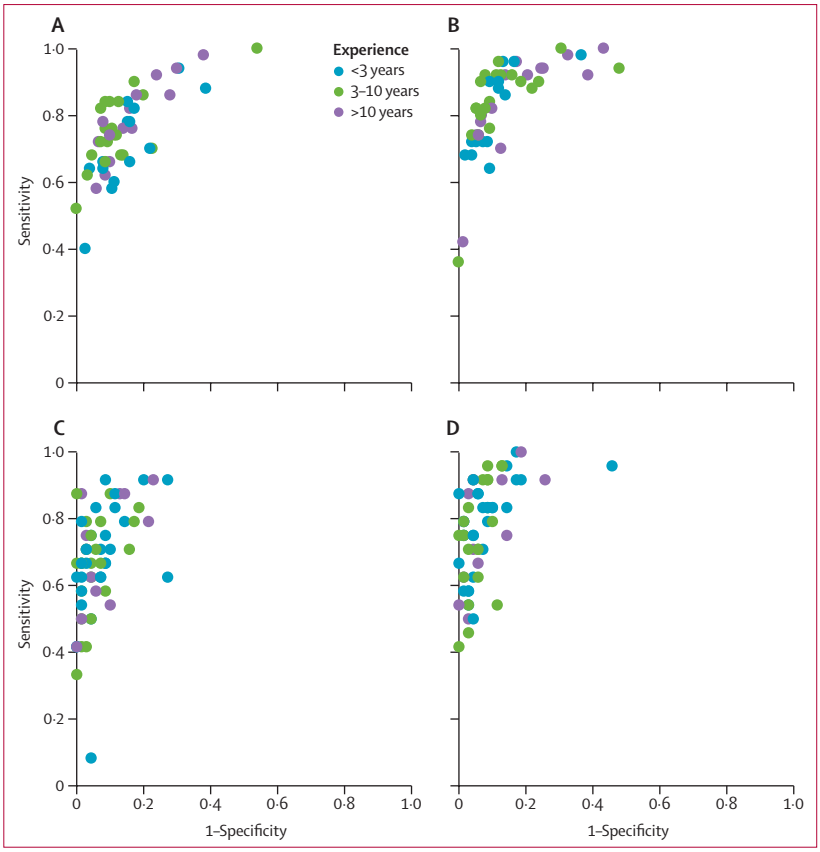

Figure 4: **Performance of general endoscopists on benchmarking test sets**
(A) General endoscopists without CADe assistance on images. (B) General endoscopists with CADe assistance on images. (C) General endoscopists without CADe assistance on video. (D) General endoscopists with CADe assistance on video. CADe=computer-aided detection.

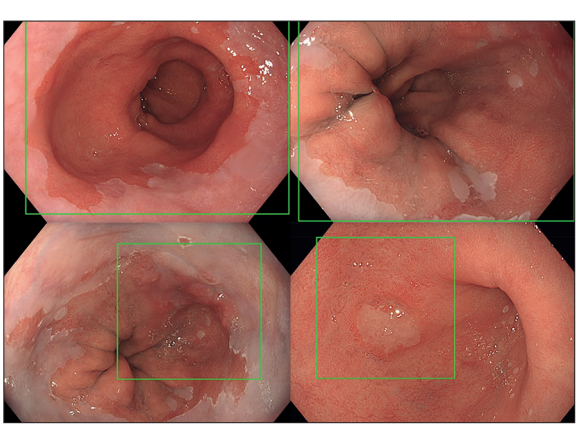

Figure 5: **Non-indicative bounding boxes and indicative bounding boxes**
The upper row shows non-indicative bounding boxes and the lower row shows indicative bounding boxes.

(ie, not all neoplastic CADe detections were accepted by the endoscopists). We think this non-acceptance of true-positive detections also explains why we could not demonstrate non-inferiority of the performance of general endoscopists with CADe assistance versus the performance of experts. These non-accepted CADe

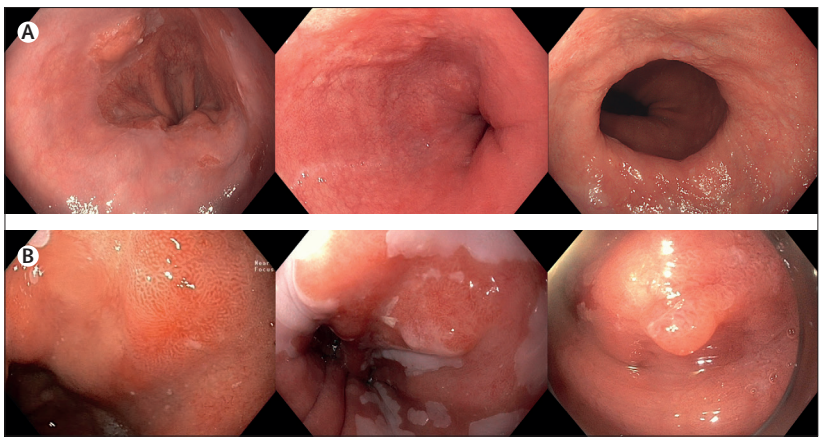

Figure 6: **Prospectively recorded images and retrospective images on neoplastic lesion**
(A) Prospectively recorded images in overview (ie, without specific focus on the neoplastic lesion). (B) Retrospective images of the neoplastic lesion.

detections might reflect detections that were interpreted as false-positive results and the ongoing process of onboarding the endoscopist with artificial intelligence (AI). However, this result might also, partly, reflect the suboptimal transfer of relevant information by the graphical user interface of the CADe system. The current graphical user interface with bounding boxes is binary in its approach, in which large or partly inaccurate bounding boxes might provide the endoscopists with confusing information compared with bounding boxes that perfectly highlight the lesion (figure 5). For videos, a briefly visible bounding box around a subtle lesion provides a different level of CADe assistance than a continuously visible bounding box. In addition, bounding boxes create an artificial outer margin of a CADe prediction and do not provide quantitative information. Future studies should focus on improving the transfer of CADe information to the endoscopist.

This study has several unique features. First, this is the largest AI study in the field of Barrett's neoplasia with an unprecedented number of images and video available for training and testing. The set-up of the BONS-AI consortium with 15 centres from seven countries ensured extensive and heterogeneous data collection, contributing to the robustness of our CADe system. In addition, 28% of the training dataset, and all images and video used for performance testing, were obtained prospectively without specific focus on any lesion—ie, mimicking the situation where lesions are often overlooked. This protocol eliminated important hidden bias that is inherent to the use of retrospectively collected imagery, in which the available neoplastic image was acquired because a lesion had been detected. Retrospectively collected images and video show neoplastic lesions in a different endoscopic configuration (ie, generally focused on the lesion) to those in the envisioned application of CADe in Barrett's surveillance (figure 6). The use of prospectively acquired images and video, obtained in overview without a specific focus on imaging lesions, ensures that images and video of non-dysplastic Barrett's oesophagus and neoplastic cases are recorded under the same circumstances. Second, we created a comprehensive ground truth for neoplasia by having 14 Barrett's oesophagus experts delineate neoplastic lesions in a standardised manner using proprietary software specifically designed for this study. Third, as well as evaluating the stand-alone performance of our CADe system, we also evaluated the interaction between the endoscopist and the CADe system in a two-phase benchmarking study with 112 general endoscopists from six countries. This approach is unprecedented in AI studies for Barrett's neoplasia. Fourth, we designed our CADe system under the specific constraints of current endoscopy systems, allowing for easy integration into the hardware platforms and real-time processing of both images and video.

This study also has some limitations. First, the CADe system was trained, validated, and tested on high-quality images and video, acquired by dedicated expert endoscopists, which could have resulted in selection bias. To improve the robustness of the CADe system, data generated by general endoscopists during standard Barrett's oesophagus surveillance should be used for training and performance testing of the CADe system. Second, the availability of subtle neoplastic imagery is relatively scarce, despite the large set-up of our BONS-AI consortium. By collecting subtle neoplastic cases for training and validation of the CADe system, it is conceivable that performance might further improve. Third, as a secondary outcome, we report the post-hoc stand-alone performance of the CADe system. It should be noted that cutoff points for stand-alone video performance of AI systems are not straightforward nor set in stone. Endoscopic videos consist of thousands of individual frames. If the detection of a single frame in a video would count as a positive detection, video-based outcomes would always result in 100% sensitivity and 0% specificity. Furthermore, brief detections of a single frame are hardly visible to the human eye. Therefore, clinically relevant detections should comprise a number of sequential positive frames over a defined amount of time. To this end, cutoff points are generally determined on the internal validation set and subsequently tested on the test set. However, this post-hoc approach is arbitrary and stand-alone results should be treated with caution. The true value of a CADe system should be evaluated based on the results of endoscopists with and without CADe assistance, which is the primary outcome in this study. Fourth, for our CADe video-based analyses, we used only a small number of post-processing steps, such as frame averaging. Several more comprehensive architectures are available to exploit the spatiotemporal information hidden in endoscopic videos, such as long short-term memory networks. Fifth, the design of the CADe system did not allow us to identify which features were used to detect

neoplasia. A key element of training deep learning systems is that the system itself decides the features that are important for correctly classifying images. Therefore, we cannot state which features (eg, colour or texture) were decisive for predicting neoplasia. Sixth, the user-preferred CADe graphical user interface should be further investigated. The current paradigm is the use of bounding boxes, which might be an obvious choice for detecting colonic polyps as they are quite distinct from the surrounding mucosa and generally have a homogeneous CADe prediction; however, for early Barrett's oesophagus neoplasia, the use of bounding boxes might not convey the optimal information on CADe predictions. Seventh, we tested our CADe system only ex vivo on images and videos, which was done for two main reasons. The first reason is that, since early Barrett's oesophagus neoplasia is relatively rare, testing CADe performance in a clinical trial would require an enormous number of patients and participating centres. Our ex vivo design allowed us to create a test set-up with both adequate statistical power and a workload that allowed many general endoscopists and Barrett's oesophagus experts to participate without compromising the quality of their assessments. The second reason is that improving endoscopists' recognition of early Barrett's oesophagus neoplasia is only useful if the CADe system is provided with a complete and adequate mucosal exposure of the Barrett's segment. The system cannot recognise lesions that are not visualised. The ex vivo design provided these circumstances when testing CADe performance. In a clinical trial, detection of early neoplasia by CADe will be compromised by incomplete visualisation of the Barrett's segment due to blind spots or incomplete cleaning of the mucosal surface. Such data will add to the background variability of the trial and further inflate the sample size and the complexity of the study. We are currently working on a quality control algorithm to improve the optimal conditions for early neoplasia detection in Barrett's oesophagus. In our opinion, clinical testing of any CADe system in endoscopy will only be successful if it operates in conjunction with such a quality control algorithm. Finally, in this study, we only included cases with visible lesions containing high-grade dysplasia or adenocarcinoma and flat-type Barrett's oesophagus with non-dysplastic histology. Future studies should evaluate the performance of our CADe system on a broader spectrum of combinations of endoscopic abnormalities and histological diagnoses.

In conclusion, this study describes the development, performance testing, and benchmarking of a CADe system for Barrett's neoplasia by the BONS-AI consortium. The CADe system outperformed general endoscopists in detecting Barrett's oesophagus neoplasia and improved their detection rate, without compromising specificity. CADe detected virtually all neoplasia in a test set of consecutive cases.

Contributors

AJdG, KNF, TGWB, JJB, MRJ, JBJ, CHJK, FvdS, and PHdW conceptualised the study. KNF, TGWB, AA, LAH, FB-S, MB, TB, LCD, PE, JJB, MRJ, JBJ, MHMGH, RCM-H, GMV, WBN, JOF-S, FvdS, OP, REP, JAvdP, KR, PS, SS, JW, BLAMW, and PHdW participated in data acquisition. KNF, MRJ, and JBJ participated in data curation and verification. KNF, MRJ, JBJ, TGWB, CHJK, AJdG, and JJB had access to raw data. KNF, TGWB, AJdG, MRJ, JBJ, SNvM, NSMM, and FvdS participated in formal statistical analysis. AJdG, JJB, FvdS, and PHdW participated in funding acquisition. KNF, TGWB, AJdG, JJB, MR, JBJ, FvdS, and PHdW designed the methodology. KNF, MRJ, and JBJ participated in project administration. TGWB, CHJK, JAvdP, and FvdS participated in software design. AJdG, JJB, FvdS, and PHdW supervised the project. KNF, AJdG, MRJ, JBJ, JJB, and FvdS participated in data validation. KNF, AJdG, JJB, MRJ, JBJ, and FvdS wrote the original draft. AA, LAH, FB-S, MB, TB, TGWB, LCD, PE, MHMGH, CHJK, RCM-H, NSMM, GMV, SNvM, WBN, JOF-S, OP, REP, JAvdP, KR, PS, SS, JW, BLAMW, and PHdW participated in reviewing and editing. AJdG, KNF, TGWB, JJB, MRJ, JBJ, CHJK, FvdS, and PHdW had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

MB is a consultant for Medtronic and a board member for Norgine and Ambu, and reports funding from Pentax Medical and medical training from Olympus. JJB reports financial support for institutional review board-approved research from C2Therapeutics and Pentax Medical, Medtronic, and Aqua Medical. TB reports consulting fees from Olympus, Boston Scientific, and Microtech; lecture fees from Olympus, Fujifilm, Pentax, Microtech, ERBE, and Medtronic; payment for expert testimony from Olympus; and participation on a data safety monitoring board or advisory board at Olympus. RCM-H reports consultant fees from Janssen. OP reports speaker honorarium from Boston Scientific, Medtronic, Fujifilm, Olympus, Aohua, Falk, and BMS. REP reports consulting fees from Medtronic and Microtech and speaker fees from Pentax. KR reports consulting fees from Olympus and lecture fees from Olympus. BLAMW reports financial support from Pentax Medical and St Antonius Research Fund; research support from Aqua Medical; consulting fees from Pentax Medical; and speaker fees from Pentax Medical and is chair of the European Society of Gastrointestinal Endoscopy guideline working group for the revision of the Barrett's oesophagus guideline. All other authors declare no competing interests.

Data sharing

De-identified participant results from the benchmarking study will be made available on reasonable request. Requests for access can be directed to m.jong3@amsterdamumc.nl. Images and code used for training and testing the computer-aided detection system will not be made available to others due to intellectual property-related constraints.

References

1 Schölvinck DW, van der Meulen K, Bergman JJGHM, Weusten BLAM. Detection of lesions in dysplastic Barrett's esophagus by community and expert endoscopists. *Endoscopy* 2017; **49:** 113–20.

2 Bergman JJGHM, de Groof AJ, Pech O, et al. An interactive web-based educational tool improves detection and delineation of Barrett's esophagus-related neoplasia. *Gastroenterology* 2019; **156:** 1299–1308.e3.

3 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:** 528–37.

4 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:** 1264–71.e1.

5 Iwagami H, Ishihara R, Aoyama K, et al. Artificial intelligence for the detection of esophageal and esophagogastric junctional adenocarcinoma. *J Gastroenterol Hepatol* 2021; **36:** 131–36.

6 Ebigbo A, Mendel R, Probst A, et al. Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. *Gut* 2019; **68:** 1143–45.

7 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:** 915–29.e4.

8 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 non-randomized trial. *Gastrointest Endosc* 2023; **97:** 422–34.

9 Ebigbo A, Mendel R, Probst A, et al. Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. *Gut* 2019; **68:** 1143–45.

10 Ghatwary N, Zolgharni M, Ye X. Early esophageal adenocarcinoma detection using deep learning methods. *Int J Comput Assist Radiol Surg* 2019; **14:** 611–21.

11 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:** 1242–50.

12 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:** 615–16.

13 van der Sommen F, Zinger S, Schoon EJ, de With PHN. Supportive automatic annotation of early esophageal cancer using local gabor and color features. *Neurocomputing* 2014; **144:** 92–106.

14 van der Sommen F, Zinger S, Curvers WL, et al. Computer-aided detection of early neoplastic lesions in Barrett's esophagus. *Endoscopy* 2016; **48:** 617–24.

15 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:** 324–36.

16 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition; June 20–25, 2009.

17 Fockens K, de Groof J, van der Putten J, et al. Linked color imaging improves identification of early gastric cancer lesions by expert and non-expert endoscopists. *Surg Endosc* 2022; **36:** 8316–25.

18 de Groof AJ, Fockens KN, Struyvenberg MR, et al. Blue-light imaging and linked-color imaging improve visualization of Barrett's neoplasia by nonexpert endoscopists. *Gastrointest Endosc* 2020; **91:** 1050–57.

19 Liu R. Higher accuracy on vision models with EfficientNet-Lite. https://blog.tensorflow.org/2020/03/higher-accuracy-on-vision-models-with-efficientnet-lite.html (accessed May 1, 2022).

20 Sandler M, Howard A, Zhu M, et al. MobileNetV2: inverted residuals and linear bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 18–23, 2018.