Endoscopy

Direct comparison of multiple computer-aided polyp detection systems

Joel Troya, Boban Sudarevic, Adrian Krenzer, Michael Banck, Markus Brand, Benjamin M Walter, Frank Puppe, Wolfram G Zoller, Alexander Meining, Alexander Hann.

Affiliations below.

DOI: 10.1055/a-2147-0571

Please cite this article as: Troya J, Sudarevic B, Krenzer A et al. Direct comparison of multiple computer-aided polyp detection systems. Endoscopy 2023. doi: 10.1055/a-2147-0571

Conflict of Interest: Prof. Alexander Meining: Consultant fees from Ovesco GmbH and Olympus Corporation. The remaining Authors declare that there is no conflict of interest.

This study was supported by Forum Gesundheitsstandort, 5409.0-001.01/15, Bayerisches Zentrum für Krebsforschung, Interdisziplinäres Zentrum für Klinische Forschung, Universitätsklinikum Würzburg (http://dx.doi.org/10.13039/501100009379), F-406

Abstract:
Background and study aims: Artificial intelligence based computer-aided polyp detection (CADe) systems receive regular updates and occasionally offer customizable detection thresholds which impact their performance, but little is known about these effects. This study aimed to compare the performance of different CADe systems on the same benchmark dataset.

Methods: 101 colonoscopy videos were used as benchmark. Each video frame with a visible polyp was manually annotated with bounding boxes resulting in 129,705 polyp images. The videos were then analyzed by three different CADe systems: two versions of GI-Genius, two detection types of EndoAID and the freely-available system EndoMind. The evaluation included an extensive analysis of sensitivity and false-positive rate, among others.

Results: EndoAID (Type A), the earlier version of GI-Genius and EndoMind detected all 93 polyps. Both the later version of GI-Genius and EndoAID (Type B) missed one polyp. The mean per-frame sensitivity was of 50.63% and 67.85% for the earlier and the latest version of GI-Genius, 65.60% and 52.95% for EndoAID (Type A and B), and 60.22% for EndoMind.

Conclusions: This study compares the performance of different CADe systems, different updates, and different configuration modes. This might help clinicians to select the most appropriate system for their specific needs.

Corresponding Author:
Joel Troya, Universitätsklinikum Würzburg Medizinische Klinik und Poliklinik II, Gastroenterology, Obermüller Straße 6, 97080 Würzburg, Germany, TroyaSebas_J@ukw.de

Affiliations:
Joel Troya, Universitätsklinikum Würzburg Medizinische Klinik und Poliklinik II, Gastroenterology, Würzburg, Germany
Boban Sudarevic, Universitätsklinikum Würzburg Medizinische Klinik und Poliklinik II, Gastroenterology, Würzburg, Germany
Adrian Krenzer, Julius Maximilians University Würzburg, Artificial Intelligence and Knowledge Systems, Würzburg, Germany
 [...]Alexander Hann, Universitätsklinikum Würzburg Medizinische Klinik und Poliklinik II, Gastroenterology, Würzburg, Germany
Direct comparison of multiple computer-aided polyp detection systems

Joel Troya¹, Boban Sudarevic¹, Adrian Krenzer², Michael Banck², Markus Brand¹, Benjamin Walter³, Frank Puppe³, Wolfram G. Zoller⁴, Alexander Meining¹, Alexander Hann¹

1. Interventional and Experimental Endoscopy (InExEn), Department of Internal Medicine II, University Hospital Würzburg, Würzburg, Germany
2. Artificial Intelligence and Knowledge Systems, Institute for Computer Science, Julius-Maximilians-Universität, Würzburg, Germany
3. Department of Internal Medicine I, University Hospital Ulm, Ulm
4. Department of Internal Medicine and Gastroenterology, Katharinenhospital, Stuttgart, Germany

Corresponding author:

M.Sc. Joel Troya
Universitätsklinikum Würzburg
Medizinische Klinik und Poliklinik II
Oberdürrbacher Str. 6, 97080 Würzburg
Germany
Tel.: +49 931 201 40040
Email: TroyaSebas_J@ukw.de

Word count: 1997

Author contributions
JT, AM and AH: Study concept and design. JT: statistical analysis. JT, FP, WZ, AM and AH: Interpretation of results, and drafting of the manuscript. JT, BS, AK, MBanck, MBrand, and BW: acquisition of data. All authors: Critical revision of the article for important intellectual content and final approval of the article.
ABSTRACT

Background and study aims: Artificial intelligence based computer-aided polyp detection (CADe) systems receive regular updates and occasionally offer customizable detection thresholds which impact their performance, but little is known about these effects. This study aimed to compare the performance of different CADe systems on the same benchmark dataset.

Methods: 101 colonoscopy videos were used as benchmark. Each video frame with a visible polyp was manually annotated with bounding boxes resulting in 129,705 polyp images. The videos were then analyzed by three different CADe systems: two versions of GI-Genius, two detection types of EndoAID and the freely-available system EndoMind. The evaluation included an extensive analysis of sensitivity and false-positive rate, among others.

Results: EndoAID (Type A), the earlier version of GI-Genius and EndoMind detected all 93 polyps. Both the later version of GI-Genius and EndoAID (Type B) missed one polyp. The mean per-frame sensitivity was of 50.63% and 67.85% for the earlier and the latest version of GI-Genius, 65.60% and 52.95% for EndoAID (Type A and B), and 60.22% for EndoMind.

Conclusions: This study compares the performance of different CADe systems, different updates, and different configuration modes. This might help clinicians to select the most appropriate system for their specific needs.

KEYWORDS

Artificial intelligence, CADe, Colonoscopy, Deep Learning, Screening
INTRODUCTION

Screening colonoscopy is considered an effective prevention measure to decrease colorectal cancer. However, 17%-48% of adenomas have been reported to be missed during this procedure [1, 2]. Computer-aided polyp detection (CADe) systems are intended to work as an adjunct to the endoscopist and help them to identify polyps. The first CADe system approved in Europe for commercial distribution was GI-Genius (Medtronic Inc., Ireland) in 2019 [3]. Since then, there has been a growing interest in demonstrating the efficiency of this type of devices. Prospective randomized controlled studies present an increase in the adenoma detection rate when endoscopist use CADe systems [4-13].

CADe systems are developed by training neural networks, usually with previously annotated images. A properly trained model can predict an output using new data. However, the output cannot be predicted in all scenarios. Therefore, it is essential to know the nature of the training data to know what they are capable of. Unfortunately, CADe manufacturers do not provide any information about this and/or how their algorithm was developed. In addition, each CADe system has been validated with different data, hindering the comparison of their performance. Furthermore, developers update CADe’s software causing an impact on their performance. For all these, it is necessary to continuously generate performance comparison studies in which systems and updates are validated with the same data.

In this study, we compare the performance of the early version (software current in March 2020, version 1.0), from now on called “first version”, and a subsequent version (software current in October 2021, version 2.0.1), from now on called “second version”, of GI-Genius; EndoAID (Olympus Medical Systems Corp., Japan, software current in March 2022) in both its
detection types A and B; and the freely-available system, EndoMind. Our aim is to compare
the sensitivity of the systems in a fully-annotated dataset to characterize their detection
strength. A comparison of other metrics such as false-positive (FP) rate and first detection
time (FDT) are performed. In addition, intersection over union (IoU) is calculated to evaluate
the accuracy of the bounding boxes.

METHODS

Study design and dataset description
A total of 244 colonoscopy videos from different patients were recorded in the University
Hospitals of Würzburg and Ulm (Germany). The videos were recorded between March 2019
and April 2020 in high-definition video signal from the endoscopy processor (Olympus CV-
190, Olympus Corporation, Japan). Details of ethics committee approval can be found in the
Supplementary Material.

Raw video analysis
A board-certified gastroenterologist and experienced endoscopist with over 4,000 performed
colonoscopies screened all the videos as described before [14]. The inclusion criteria were
examinations performed for screening purposes or post polypectomy surveillance. Exclusion
criteria are described in the Supplementary Material and Supplementary figure 1. Using a
custom-made annotation tool, the colonoscopies were analyzed in a deep frame-by-frame
process and bounding boxes were drawn around each frame containing a polyp [15]. Details
about the annotation process can be found in Supplementary Material and Supplementary
Figure 2.
CADe data obtention

All the raw colonoscopy videos were processed by each CADe systems in the same way. A video converter was used to send the signal from a laptop to the different CADe systems (Mini Converter UpDownCross HD, Blackmagic Design Pty. Ltd., Australia). A video recorder (DeckLink Mini Recorder, Blackmagic Design Pty. Ltd., Australia) was used to record the HD-signal of each CADe system. A custom algorithm to detect the bounding box locations was developed using Python (Python Software Foundation, version 3.8) (Supplementary Material).

Outcomes

The primary outcome measure of the study was sensitivity. Per-polyp sensitivity was defined as the ratio between the polyp detected in at least one frame by the CADe, and the total number of polyps. The per-frame sensitivity was defined as the ratio between the number of correct identified images with polyps and the total number of images of that particular polyp. In this case the duration that the polyp appears in the image is accounted. Multiple metrics have been included as secondary outcomes: IoU, FDT and FP rate. IoU measures the accuracy of the predicted bounding boxes by comparing the overlap with a ground-truth bounding box (Supplementary Figure 3). More details of these metrics and the statistical analysis can be found in Supplementary Material.

RESULTS

Baseline Characteristics

Out of 244 recorded routine colonoscopies, 143 colonoscopies met the exclusion criteria. Thus, a total of 101 colonoscopy videos were used to analyze the performance of each of the CADe systems. From a total of 2161818 image frames, 464186 were considered part of a
polypectomy, 56902 were acquired with Narrow Band Imaging (NBI) light, 37445 frames were repeated image frames due to freezes for documentation and 97105 consisted of images of the rectum. A total of 45 (44.55%) videos contained at least one polyp. The total number of polyps was 93. This accounted for 129705 (8.61%) image frames (Supplementary figure 1). In total, 1506180 image frames were processed by each system, resulting in a dataset of 7530900 images.

The patients and polyp characteristics with the accompanying histology are presented in Supplementary Table 1.

**Primary outcome**

**Sensitivity**

Per-polyp sensitivity was 100% for the first version of GI-Genius, EndoAID in detection Type A and EndoMind. Both the second version of GI-Genius and EndoAID in Type B missed 1 polyp (Figure 1). The second version of GI-Genius did not detect a sessile serrated adenoma (SSA) of Paris 0-IIa located in the right colon that was present for 10.20 seconds. This would have had a 7-year delayed impact on patient follow-up according to German and U.S. guidelines [16, 17]. In the case of EndoAID (Type B), it did not detect a polyp in the right colon that was also not detected by the endoscopist of Paris 0-IIa and present for 0.87 seconds. In this case, there would not be any delay in patient follow-up. The overall mean per-frame sensitivity for each system was of 50.63% (CI: 45.20-56.07%) for the first version of GI-Genius, 67.85% (CI: 63.26-72.43%) for the second version of GI-Genius, 65.60% (CI: 60.26-70.95%) for EndoAID (Type A), 52.95% (CI: 46.92-58.99%) for EndoAID (Type B), and 60.22% (CI: 54.66-65.78%) for EndoMind (Table 1). The median per-frame sensitivity was significantly different between all the devices except between the second version of GI-Genius and EndoAID in Type A (p=0.460), and the first version of GI-Genius and Olympus in Type B (p=0.242). The median
per-frame sensitivity was significantly lower for flat polyps (51.70%, IQR: 29.35-72.58) when compared to 0-Ip (85.90%, IQR: 71.40-95.40) or 0-Is (81.00%, IQR: 64.25-89.15) morphology in all the devices.

**Secondary outcomes**

*First detection time*

The mean FDT for each system was of 1510 (CI: 1125-1895) ms for the first version of GI-Genius, 607 (CI: 411-803) ms for the second version of GI-Genius, 659 (CI: 410-909) ms for EndoAID (Type A), 1316 (CI: 951-1682) ms for EndoAID (Type B), and 1083 (CI: 627-1539) ms for EndoMind (**Supplementary Table 2**). The median FDT was significant different between all the systems. All the systems presented longer median FDT for polyps with 0-IIa morphology (350ms, IQR: 167-1442ms) when compared to 0-Ip (333ms, IQR: 133-533ms, p=0.063) and to 0-Is (233ms, , IQR: 133-583ms, p=0.002).

*Intersection over union*

The mean IoU was of 58.18 (CI: 58.0-58.36)% for the first version of GI-Genius and 61.06 (CI: 60.91-61.21)% for the second version of GI-Genius. The mean IoU of EndoAID was of 63.54 (CI: 63.38-63.70)% in Type A and of 66.13 (CI: 65.98-66.29)% in Type B. Regarding EndoMind, the value for the mean IoU was of 68.32 (CI: 68.15-68.48)% (**Figure 2**). All IoU distribution mean values have been tested significantly different from each other.

*False positive rate*

The first version of GI-Genius presented a total of 41411 FP image frames, equivalent to a rate of 2.75% of all the images. The second version of the device showed a total of 57278 FP images (3.80%). Regarding EndoAID, in detection Type A there were 38012 FP images (2.52%) and in Type B there were 9432 FP images (0.63%). The freely available system EndoMind presented 55631 FP images accounting for a FP rate of 3.69% (**Figure 3**).
A summary table for the results as well as additional metrics such as per-box mean precision of each of the systems are shown in Table 2. Video 1 shows the different devices during the recognition of an adenoma.

DISCUSSION

Recently published randomized clinical trials present evidence of the ability of CADe systems like GI-Genius and EndoAID in detecting more adenomas in comparison to examinations without CADe [9, 13, 18]. Since then, CADe systems quickly expanded in clinical practice [11, 12, 19, 20]. However, as already discussed, it is difficult to compare the CADe systems between each other to establish which performs better. Additionally, updates of existing systems might influence their ability to detect polyps and thus impact performance. For these reasons, CADe systems of different manufacturers and of different versions need to be compared along time using the same dataset under the same conditions.

In this study we compared the evolution of the GI-Genius system. This was to our knowledge never described before. The first version might be closer or equal to the one that was used in Repici et al. in the study period of September to November 2019 [9]. The second might resemble the one that was used also in Repici et al. in a study period of February to December 2020 [18]. We have identified that the later version is significantly more sensitive than the first version and needs less time to detect polyps. However, the number of FPs is also significantly higher.

Customization of systems will be implemented more and more [21]. In this sense, EndoAID uses two different detection types that are described in the manual as: Type A, detects more
potential colorectal polyps than Type B; Type B, tends to suppress more false detections than Type A. In Schauer et al. and Gimeno-García et al., the detection Type A was used [13, 22]. In our study we could confirm the high sensitivity reported in all the studies analyzing the CADe system. On the other hand, detection Type B was not able to detect one polyp. However, the number of FPs was significantly reduced, leading to high specificity.

In the previous study published by our group, EndoMind had a significant higher number of FPs and a significant lower FDT when compared to our current work [23]. One reason might be that the hardware of EndoMind, which has been used in the present study predicts every third frame rather than single frames to reduce the number of FPs and having less delay in the image processing pipeline. Instead, in our previous work the neuronal network EndoMind predicted in each single frame.

This study has some limitations. EndoAID system uses the EVIS X1 CV-1500 videoprocessor, therefore videocolonoscopes of the Serie 1500 would be a great option. Anyhow, CF-HQ190 videocolonoscopes have the great advantage of being supported by all the CADe systems compared in this study, including EndoAID. Thus, we excluded from our dataset the videos recorded with CF-H180 videocolonoscopes. Considering that this was a retrospective study, not all detected polyps by the endoscopist were resected and therefore, histology of some cases is not available. Finally, while our study provides valuable insights into the performance trends of different CADe systems, the retrospective and exploratory nature of the analysis limits the comparison.
In summary, our present study describes for the first time the performance of three artificial intelligence polyp detection systems in a same dataset. In addition, frame-by-frame analysis gives much more robust results and a clearer picture of how the systems perform in real conditions. It has been observed that the GI-Genius software update significantly increases sensitivity, and the impact in the behavior of the EndoAID system depending on the detection type chosen. Finally, EndoMind, a freely available system developed in a public hospital, has been shown to perform similarly to commercially available systems. Based on the outcomes here presented, clinicians might decide which CADe system best suits their needs by selecting the one with the preferred sensitivity-specificity balance.

CONFLICT OF INTERESTS
Author Alexander Meining is a paid consultant for Ovesco GmbH and Olympus Corporation for unrelated research. The remaining Authors declare that there is no conflict of interest.

FUNDING INFORMATION
Wolfram G. Zoller and Alexander Hann receive funding from the state government of Baden-Württemberg, Germany (Funding cluster “Forum Gesundheitsstandort Baden-Württemberg”) to research and develop artificial intelligence applications for polyp detection in screening colonoscopy (funding number 5409.0-001.01/15). Alexander Meining and Frank Puppe receive funding from the IZKF Würzburg (funding number F-406) and the Bavarian Center for Cancer Research (BZKF) for further implementation and development of artificial intelligence for detection of (pre-) neoplastic lesions.

REFERENCES


FIGURE LEGENDS

Figure 1. Images of polyps that were not detected by GI Genius version 2 (left) or EndoAID in detection Type B (right).

Figure 2. Violin plot showing the distribution of the intersection over union between the ground-truth manually annotated bounding box and the displayed CADe bounding box. It is similar to a box plot, with the addition of a rotated kernel density plot on each side. Additionally, a box plot is found inside. The ends of the box represent the lower and upper quartiles, while the median (second quartile) is marked by a continuous line inside the box. The discontinued line represents the mean value of the intersection over union distribution.

Figure 3. Bar plot showing the total amount of false-positive image frames triggered by every CADe system analyzed.

VIDEO LEGENDS

Video 1. Visualization of the CADe systems compared during the detection of an adenoma. Above from left to right: Goldstandard, GI Genius v1, GI Genius v2. Below from left to right: EndoAID A, EndoAID B, Endomin.

Video image. Representative image of Video 1.
TABLE LEGENDS

Table 1. Sensitivity for each CADe system analyzed.

Table 2. Summary table of the results.

Supplementary Table 1. Characteristics of the patients and the polyps included in the dataset.

Supplementary Table 2. First detection time for each CADe system analyzed.

SUPPORTING INFORMATION LEGENDS

Supplementary Figure 1. Flow diagram of data selection. IBD, inflammatory bowel disease; SSP, serrated polyposis syndrome; ESD, endoscopic submucosal dissection; EMR, endoscopic mucosal resection; NBI, narrow-band imaging.

Supplementary Figure 2. Screenshot of the custom review program developed to check the ground-truth annotations.

Supplementary Figure 3. Explanation of the intersection over union metric and the criteria used to classify the polyp-containing image frames into true or false according to the intersection over union value.
Supplementary material

Direct comparison of multiple computer-aided polyp detection systems

Authors: Joel Troya, Boban Sudarevic, Adrian Krenzer, Michael Banck, Markus Brand, Benjamin Walter, Frank Puppe, Wolfram G. Zoller, Alexander Meining, Alexander Hann

Study flow diagram

Supplementary Figure 1 shows the study flow diagram of the colonoscopy videos included and excluded in the study. Exclusion criteria included inflammatory bowel disease, bleeding, poor bowel preparation, incomplete colonoscopy, stenosis, defective videos, serrated polyposis syndrome, endoscopic submucosal dissection clip, hemicolecotomy, full-thickness resection in the rectum, lymphomas or tumors, xanthomatosis, Kaposi sarcoma and endoscopic mucosal resection.

All the colonoscopies were performed with CF-HQ190AL colonoscopes (Olympus Corporation, Japan). A standard split-dose regimen was used to prepare all the patients for the colonoscopy.

Annotation methodology

A custom-made annotation tool was used to annotate in a deep frame-by-frame process all the included colonoscopies. A board-certified gastroenterologist was in charge of this process. The annotated labels included all the frames with polyps and the establishing of the withdrawal and resection times. In addition, the size and morphology of the detected polyps were annotated at this point. To exclude the presence of multiple hyperplastic polyps in the rectum, this region was excluded from the analysis. Subsequently in this study, two trained medical students drew bounding boxes in each of the frames that contained a polyp to obtain its precise location in the image and setting the ground-truth. All these bounding boxes were then checked an additional time by a second board-certified gastroenterologist and corrections were manually performed if necessary. In addition, the hole annotation process was supervised by an experienced biomedical engineer.

To perform the annotations and revisions of the frames, a custom-software revision tool was developed using the library PyQt5 in Python 3.8 (Supplementary figure 2). The revision tool has a main screen where the image is displayed. The tasks or images to review are listed in a scroll tab in the right of the screen. The ground-truth bounding boxes previously annotated are displayed in red and can be reset and corrected. The speed of image visualization could be chosen with a slider. If the polyp was not visible in the screen, the endoscopist could press the button “Where is the polyp?” which would generate the label “Polyp not identified” in the dataset. False positive detections were also screened to discard the presence of additional lesions. In comparison to the previously published...
analysis [1], we included only white light images since the Olympus EndoAID predicts only in this mode. We also excluded all endoscopies performed with a colonoscope CF-H180 since EndoAID does not support this endoscope type.

**Bounding box detection and localization**

In order to detect the presence of bounding boxes displayed by a computer-aided polyp detection (CADe) system in an image frame, two different approaches were used. On the one hand, classical image transformation methods allowed us to detect and locate the bounding boxes. On the other hand, a neural network trained to detect the bounding boxes was developed. Both algorithms ran in parallel. When there were discrepancies between them, the user was deciding if there was presence of bounding box or not and the location. With this methodology, we ensure that all the bounding boxes were correctly determined.

The following small sections show the preprocessing and detection algorithms used on the “classical” approach and details about the developed neural network trained to detect the bounding boxes.

**Preprocessing**

```python
import numpy as np
import cv2
from skimage import morphology
from skimage.morphology import square

# Crop the image
image = image[:, 320:1410, :]

# Set minimum and maximum HSV value thresholds
lower = np.array([22, 144, 95])
upper = np.array([62, 246, 255])

# Create HSV image and create a mask using the thresholds
hsv = cv2.cvtColor(image, cv2.COLOR_BGR2HSV)
mask = cv2.inRange(hsv, lower, upper)

# Slice the hsv
imask = mask>0
image = np.zeros_like(hsv, np.uint8)
image[imask] = hsv[imask]

# Separate by channel
b, image, r = cv2.split(image)

# Blur the image
image = cv2.bilateralFilter(image, 9, 25, 25)

# Sharpen the image
kernel = np.array([[0, -1, 0], [-1, 5, -1], [0, -1, 0]])
image = cv2.filter2D(src=blurred, ddepth=-1, kernel=kernel)
```
# Apply dilation
image = morphology.dilation(image, square(3))

# Apply binary threshold
(thresh, image) = cv2.threshold(res, 128, 255, cv2.THRESH_BINARY | cv2.THRESH_OTSU)

# Apply 3x dilation
for counter in range(3):
    image = morphology.dilation(image, square(3))

return image

Detecting boxes

import cv2
import numpy as np

# Find contours
contours, hierarchy = cv2.findContours(image, cv2.RETR_TREE, cv2.CHAIN_APPROX_SIMPLE)

# If the contour area is bigger than a threshold, then it is probably a bounding box
bboxes = []
for contour in contours:
    if cv2.contourArea(contour) > 800:
        bboxes.append(cv2.boundingRect(contour))

# Check if the bounding boxes are inside bounding boxes
bboxes = is_inside(bboxes)

# Post-process boxes using intersections, overlapping and location in the image
bboxes = process_coordinates(bboxes)

Artificial intelligence to detect boxes

For model training we used the pretrained ResNet50 based on deep learning and provided by the tensorflow.keras.applications library. The model was trained with 6575 positive bounding-box images and 6314 negative bounding-box images in a batch size of 29 for 5 epochs. Binary crossentropy from the keras library was used as a loss function and a learning rate of 0.0001 was set for the Adam optimizer. The model was trained with a NVIDIA RTX3080 graphic processing unit.

Secondary outcomes

The secondary outcomes analyzed in this study are: the intersection over the union (IoU), the first detection time (FDT) and the false-positive rate. All of them have been analyzed in a frame-by-frame manner.
The IoU is a standard metric used to measure the accuracy of object detection by comparing the overlap between a predicted bounding box and a ground-truth bounding box with the total area of both. The IoU was calculated for each displayed bounding box of each CADe system and was used to discriminate between correct and incorrect bounding boxes and calculate each CADe system’s sensitivity (Supplementary figure 3). A bounding box with IoU>0 was regarded as correct (TP). This is equivalent to considering the activation correct if there is an intersection between the gold-standard and the predicted bounding box. As analyzed by Tran et al., this approach might better approximate the clinical need [2]. The FDT was another secondary outcome. FDT is defined as the minimum time required for the CADe to detect a polyp, without considering the latency of the system. Lastly, the FP rate was analyzed as the proportion between the number of non-polyp containing images in which the CADe falsely predicted a polyp and the total number of frames.

Statistical analyses

To test if the data follows a Gaussian distribution, the Shapiro-Wilk test was used. To test if there are significant differences in the per-polyp sensitivity and in the FDT distributions across the different CADe systems, the Mann-Whitney U test was used. To test if the mean values of the IoU distributions were significantly different, the t-test was used. The point estimates for the mean per-frame sensitivity, the mean FDT, and the mean IoU are presented with a 95% confidence interval. The point estimates for the median per-frame sensitivity for polyp morphology are presented with the interquartile range. All statistics were performed using SciPy package in Python version 3.8 [3].

Ethics

The retrospective analysis of data was reviewed and approved by the Ethics Committee of the University Hospital Würzburg, approval number 2021032901. Patients were not required to give informed consent for this retrospective analysis.

References

**Supplementary Table 1.** Characteristics of the patients and the polyps included in the dataset.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
</tr>
<tr>
<td>Male, n (%)</td>
<td>47 (46.53)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>54 (53.47)</td>
</tr>
<tr>
<td><strong>Age, mean (range)</strong></td>
<td>60 (19-89)</td>
</tr>
<tr>
<td><strong>BBPS, mean (range)</strong></td>
<td>7.45 (6-9)</td>
</tr>
<tr>
<td><strong>Colonoscopy purpose</strong></td>
<td></td>
</tr>
<tr>
<td>Screening, n(%)</td>
<td>27 (26.73)</td>
</tr>
<tr>
<td>Symptomatic, n(%)</td>
<td>74 (73.27)</td>
</tr>
<tr>
<td><strong>Polyps, n</strong></td>
<td>93</td>
</tr>
<tr>
<td><strong>Polyps per patient, mean (range)</strong></td>
<td>0.92 (0-6)</td>
</tr>
<tr>
<td><strong>Polyp detection rate, (%)</strong></td>
<td>44.55</td>
</tr>
<tr>
<td><strong>Adenoma detection rate, (%)</strong></td>
<td>32.67</td>
</tr>
<tr>
<td><strong>Histology, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Adenoma</td>
<td>39 (41.94)</td>
</tr>
<tr>
<td>Hyperplastic</td>
<td>13 (13.98)</td>
</tr>
<tr>
<td>SSA</td>
<td>17 (18.28)</td>
</tr>
<tr>
<td>Other</td>
<td>24 (25.81)</td>
</tr>
<tr>
<td><strong>Paris classification, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>0-Ip</td>
<td>5 (5.38)</td>
</tr>
<tr>
<td>0-Is</td>
<td>28 (30.11)</td>
</tr>
<tr>
<td>0-IIa</td>
<td>60 (64.52)</td>
</tr>
<tr>
<td><strong>Size, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt; 5 mm</td>
<td>54 (58.06)</td>
</tr>
<tr>
<td>5 – 10 mm</td>
<td>24 (25.81)</td>
</tr>
<tr>
<td>&gt; 10 mm</td>
<td>15 (16.13)</td>
</tr>
<tr>
<td><strong>Location, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Right colon</td>
<td>50 (53.76)</td>
</tr>
<tr>
<td>Left colon</td>
<td>28 (30.11)</td>
</tr>
<tr>
<td>Rectum</td>
<td>15 (16.13)</td>
</tr>
</tbody>
</table>

SSA, Sessile serrated adenoma
<table>
<thead>
<tr>
<th>Overall</th>
<th>Paris Classification</th>
<th>Size</th>
<th>Location</th>
<th>Histology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-Ip</td>
<td>0-Is</td>
<td>0-IIa</td>
<td>&lt;5 mm</td>
</tr>
<tr>
<td>GI Genius version 1</td>
<td>1510 (1125-1895)</td>
<td>520 (216-824)</td>
<td>913 (493-1333)</td>
<td>1872 (1325-2419)</td>
</tr>
<tr>
<td>GI Genius version 2</td>
<td>607 (411-803)</td>
<td>453 (0-1348)</td>
<td>571 (268-875)</td>
<td>637 (366-909)</td>
</tr>
<tr>
<td>EndoAID Type A</td>
<td>659 (410-909)</td>
<td>387 (0-813)</td>
<td>287 (141-433)</td>
<td>856 (484-1229)</td>
</tr>
<tr>
<td>EndoAID Type B</td>
<td>1316 (951-1682)</td>
<td>960 (0-2650)</td>
<td>699 (314-1083)</td>
<td>1640 (1121-2158)</td>
</tr>
<tr>
<td>EndoMind</td>
<td>1083 (627-1539)</td>
<td>167 (0-340)</td>
<td>376 (193-559)</td>
<td>1489 (808-2171)</td>
</tr>
</tbody>
</table>

**Supplementary Table 2.** First detection time for each CADe system analyzed.

ms: milliseconds; HP: Hyperplastic polyp; SSL: Sessile serrated lesion
<table>
<thead>
<tr>
<th>CADe System</th>
<th>Per-polyp sensitivity (%)</th>
<th>Paris Classification</th>
<th>Size</th>
<th>Location</th>
<th>Histology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>0-Ip</td>
<td>0-Is</td>
<td>0-IIa</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;5 mm</td>
<td>5-10 mm</td>
<td>&gt;10 mm</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Right colon</td>
<td>Left colon</td>
<td>Rectum</td>
</tr>
<tr>
<td>GI Genius version 1</td>
<td>100</td>
<td>50.63 (45 - 56)</td>
<td>79.80 (56 - 100)</td>
<td>64.96 (58 - 72)</td>
<td>41.52 (35 - 48)</td>
</tr>
<tr>
<td>GI Genius version 2</td>
<td>98.92</td>
<td>67.85 (63 - 72)</td>
<td>85.36 (62 - 100)</td>
<td>79.81 (74 - 85)</td>
<td>60.80 (55 - 67)</td>
</tr>
<tr>
<td>EndoAID Type A</td>
<td>100</td>
<td>65.60 (60 - 71)</td>
<td>81.60 (59 - 100)</td>
<td>81.44 (75 - 87)</td>
<td>56.88 (50 - 64)</td>
</tr>
<tr>
<td>EndoAID Type B</td>
<td>98.92</td>
<td>52.95 (47 - 59)</td>
<td>70.16 (39 - 100)</td>
<td>72.35 (65 - 80)</td>
<td>42.47 (35 - 50)</td>
</tr>
<tr>
<td>EndoMind</td>
<td>100</td>
<td>60.22 (55 - 66)</td>
<td>73.66 (39 - 100)</td>
<td>77.37 (71 - 84)</td>
<td>51.10 (44 - 58)</td>
</tr>
</tbody>
</table>

Table 1. Sensitivity for each CADe system analyzed.  
HP: Hyperplastic polyp; SSL: Sessile serrated lesion
<table>
<thead>
<tr>
<th>Metric (mean)</th>
<th>GI Genius version 1</th>
<th>GI Genius version 2</th>
<th>EndoAID Type A</th>
<th>EndoAID Type B</th>
<th>EndoMind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per-polyp sensitivity, %</td>
<td>100</td>
<td>98.92</td>
<td>100</td>
<td>98.92</td>
<td>100</td>
</tr>
<tr>
<td>Per-frame sensitivity, %</td>
<td>50.63</td>
<td>67.85</td>
<td>65.60</td>
<td>52.95</td>
<td>60.22</td>
</tr>
<tr>
<td>FP rate, %</td>
<td>2.75</td>
<td>3.80</td>
<td>2.52</td>
<td>0.63</td>
<td>3.69</td>
</tr>
<tr>
<td>FP rate per colonoscopy, %</td>
<td>2.43</td>
<td>3.40</td>
<td>2.41</td>
<td>0.51</td>
<td>3.90</td>
</tr>
<tr>
<td>First Detection Time, ms</td>
<td>1510</td>
<td>607</td>
<td>659</td>
<td>1316</td>
<td>1083</td>
</tr>
<tr>
<td>Intersection over union, %</td>
<td>58.18</td>
<td>61.06</td>
<td>63.54</td>
<td>66.13</td>
<td>68.32</td>
</tr>
<tr>
<td>Precision, %</td>
<td>59.66</td>
<td>57.01</td>
<td>66.58</td>
<td>87.20</td>
<td>54.99</td>
</tr>
<tr>
<td>Specificity, %</td>
<td>96.94</td>
<td>95.77</td>
<td>97.19</td>
<td>99.30</td>
<td>95.89</td>
</tr>
<tr>
<td>F1 Score, %</td>
<td>59.15</td>
<td>63.91</td>
<td>69.43</td>
<td>71.77</td>
<td>59.60</td>
</tr>
</tbody>
</table>

Table 2. Summary table of the results. All the metrics except per-polyp sensitivity have been assessed in a frame-by-frame manner.

FP: False-positive
244 routine colonoscopies recorded

143 colonoscopies excluded
  - IRD: 56
  - Bleeding: 5
  - Poor Bowel Preparation: 15
  - Incomplete: 19
  - Stenosis: 4
  - Defective videos: 26
  - Colonoscopies with CPIH380: 6
  - SSP-Syndrom: 1
  - ESD clip: 1
  - Hemicolecotomy: 3
  - Full-Thickness Resection in Rectum: 1
  - Lymphoma in colon: 1
  - Tumor ascendens: 1
  - Xanthomatosi: 1
  - Kaposis sarcoma: 1
  - Tumor in sigmoid: 1
  - EMR with clip in ileum: 1

101 colonoscopies included

45 colonoscopies with at least one polyp
  - 93 polyps

2 161 818 image frames

56 colonoscopies without polyps

1 506 180 included image frames

129 705 polyp image frames
1 376 475 non-polyp image frames

655 638 frames excluded
  - Polypectomies: 464 186
  - NBI light: 56 902
  - Image frames for documentation: 37 445
  - Images in rectum: 97 105
Intersection over Union = \frac{\text{Area of overlap}}{\text{Area of union}} = \frac{\text{Goldstandard bounding box}}{\text{CADe bounding box}}

<table>
<thead>
<tr>
<th>IoU</th>
<th>Frame classified as</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>IoU &gt; 0</td>
<td>True positive</td>
<td><img src="image1" alt="True positive examples" /></td>
</tr>
<tr>
<td>IoU = 0</td>
<td>False positive</td>
<td><img src="image2" alt="False positive examples" /></td>
</tr>
</tbody>
</table>