Author's response to decision letter – Point-to-point response

Dear editor and reviewer,

Thank you for reviewing our submission to Artificial Intelligence in Gastrointestinal Endoscopy, providing us with valuable comments, and offering us the opportunity to resubmit our manuscript after a thorough revision. We have revised the paper such that all concerns of the reviewers are addressed and the suggested alterations are incorporated. We address these concerns in a point-wise fashion.

Yours sincerely,

Quirine van der Zander
also on behalf of the co-authors

Reviewer comments:
Reviewer #1:
Scientific Quality: Grade B (Very good)
Language Quality: Grade A (Priority publishing)
Conclusion: Minor revision
Specific Comments to Authors:
I congratulate the authors on this interesting important piece of work which adds important evidence to the evaluation of CADx for colorectal polyps, albeit with a small sample size. Please see my minor revisions below.

Abstract (results section) - please also report the proportion of polyps that were diagnosed with low confidence when reporting the results of the self-critical AI4CRP.
We thank the reviewer for this suggestions.

We have adjusted the following in the abstract:
"Self-critical AI4CRP, excluding 14 low confidence characterizations (27.5% [14/51]), had a diagnostic accuracy of 89.2%, sensitivity of 89.7%, and specificity of 87.5%, which was higher compared to AI4CRP."

Abstract (results section) – It needs to be clearer that the numerical increase in the endoscopist's performance was after reviewing both CADx systems (AI4CRP and CAD-Eye)
We have adjusted the following in the abstract:
"Diagnostic performances of the endoscopist alone (before AI) increased non-significantly after reviewing the CADx characterizations of both AI4CRP and CAD EYE (AI-assisted endoscopist)."

Abstract (conclusions section) – Please rephrase the final sentence to reflect that the endoscopist performance was non-significantly higher then both CADx-systems.

We carefully considered the suggested rephrasing of the final sentence in the abstract. The results on the statistical testing of the diagnostic performances of (self-critical) AI4CRP, CAD EYE, the endoscopists alone, and the AI-assisted endoscopist, can be find in Supplementary figure 2. As can be seen for the results, the significance level is not equal for all diagnostic metrics (some are significantly different and some are not):

For sensitivity, both the endoscopists alone and the AI-assisted endoscopists performed significantly higher compared to AI4CRP and CAD EYE, but not compared to self-critical AI4CRP.
For specificity, the AI-assisted endoscopists performed non-significantly higher compared to AI4CRP and self-critical AI4CRP, while CAD EYE performed non-significantly higher compared to both the endoscopists alone and the AI-assisted endoscopists.
For diagnostic accuracy, both the endoscopists alone and the AI-assisted endoscopists performed significantly higher compared to CAD EYE, but non-significantly compared to AI4CRP and self-critical AI4CRP.

Therefore, it cannot be said that the endoscopists performed non-significantly higher than both CADx-systems, since this does not account for all diagnostic metrics. We therefore added a sentence to the result section for clarification and removed the final sentence in the abstract.

Addition to the result section:
"Diagnostic performances of the AI-assisted endoscopist were higher compared to both CADx-systems, except for specificity for which CAD EYE performed best."

Introduction – please clarify that non expert endoscopist do not consistently meet quality standards set by ASGE and ESGE.
We thank the reviewer for this remark and clarified the statement:

"Despite these optimizations, endoscopists do not consistently meet quality standards set by the American society for gastrointestinal endoscopy (ASGE) and the European society of gastrointestinal endoscopy (ESGE) for implementation of the resect-and-discard and diagnose-and-leave strategies based on optical diagnosis. The first strategy entails diminutive (≤5 mm) colorectal polyps to be resected and discarded without histopathological assessment under the condition of a ≥90% agreement in the post-polypectomy surveillance interval between the optical and histopathological diagnosis. The second strategy states that diminutive hyperplastic polyps in the rectosigmoid can be left in situ if a negative predictive value (NPV) of ≥90% is reached for the optical diagnosis of adenomatous polyps."

Methodology – Are you able to expand on the additional training in optical diagnosis that the endoscopist underwent?
We thank the reviewer for the question.

We specified the additional training:
"The endoscopist was additionally trained in optical diagnosis (succeeding several training sessions in optical diagnosis by the ESGE), and performed optical diagnoses on a regular basis according to the ESGE curriculum for optical diagnosis, and is a teacher in optical diagnosis training sessions.[1]"

Methodology – please kindly expand on the sample size calculation (30 patients) which was based on a previous CADx feasibility study.
We elaborate on the sample size below:

The CADx feasibility study we refer to aimed to investigate the preliminary diagnostic performances of a CADx-system for Barrett's esophagus (both nondysplastic Barrett's esophagus and Barrett's neoplasia). This study included 20 patients with Barrett's esophagus.[18] The aim of the current study was also to evaluate the feasibility of the real time use of a CADx-system (AI4CRP). Due to the feasibility design of the study, no formal sample size calculation was performed. Therefore, the sample size was based on the mentioned study. We decided to include 30 patients. Although we also use two histopathology categories (benign and premalignant), for AI4CRP the premalignant category consisted of both tubular adenomas and sessile serrated lesions. Therefore, we decided to increase the sample size to 30.

18 de Groof AJ, Struyvenberg MR, Fockens KN, van der Putten J, van der Sommen F, Boers TG, Zinger S, Bisschops R, de With PH, Pouw RE, Curvers WL, Schoon EJ, Bergman J. Deep learning algorithm detection of Barrett's neoplasia with high accuracy during live endoscopic procedures: a pilot study (with video). Gastrointest Endosc 2020 [PMID: 31926965 DOI: 10.1016/j.gie.2019.12.048]

We added the following to the methods:
"Due to the feasibility design of the study, no formal sample size calculation was performed. The sample size (n=30 patients) was based on a previous CADx feasibility study.[18]"

Results – Are you able to report on the number of images that were excluded due to 'motion blur' and 'out of focus' images.
We thank the reviewer for this suggestion and reported the number of excluded images in the results:

"Eight images were excluded because the images were out of focus and four because of motion blur. For these colorectal polyps, a second image was taken by the endoscopist."

Results – was any quantifiable testing performed for the latency?
We thank the reviewer for the question. Latency was not measured by the system itself or by the investigators. From a previous study with AI4CRP, we know that the mean computation time per image was 0.0258 seconds (SD 0.0148 seconds).[12] Small differences in latency are therefore not noticeable for endoscopists. We therefore defined feasibility as, among others, no noticeable clinically relevant latency.

12 **van der Zander QEW**, Schreuder RM, Fonollà R, Scheeve T, van der Sommen F, Winkens B, Aepli P, Hayee B, Pischel AB, Stefanovic M, Subramaniam S, Bhandari P, de With PHN, Masclee AAM, Schoon EJ. Optical diagnosis of colorectal polyp images using a newly developed computer-aided diagnosis system (CADx) compared with intuitive optical diagnosis. Endoscopy 2021; 53(12): 1219-1226 [PMID: 33368056 DOI: 10.1055/a-1343-1597]

We have specified this in the methods:
"Feasibility was defined as seamless video output reception from the endoscopy processor without noticeable clinically relevant latency (the time from capturing the endoscopic image to outputting the analyzed results)[17] and seamless operation of the software in obtaining characterizations. Latency was not measured by the AI4CRP system itself or by the investigators since it is known from previous studies that small differences in latency were not noticeable for endoscopists, and therefore only clinically noticeable latency was deemed relevant.[12]"

Results - Please rephrase the reporting of the expert endoscopist performance to reflect that the endoscopist diagnostic performance was non-significantly higher than both CADx-systems (instead of "did not increase significantly", as the study was not powered to detect this increase).
We agree with the reviewer that the study was not powered to detect differences in endoscopists performances before and after CADx reviewing. We have added this in the result section.

However, as stated above, the results of the endoscopists compared to the two CADx systems are not equal regarding significance. Therefore, it cannot be said that the endoscopists performed non-significantly higher than both CADx-systems, as is shown in the last part of the result section ("Comparing diagnostic performances …").

We rephrased the sentence in the Results:
"Although this study was not powered to detect a difference between the endoscopist alone and the AI-assisted endoscopist, after reviewing characterizations of both CADx-systems specificity, PPV, NPV, and diagnostic accuracy increased non-significantly for the AI-assisted endoscopist (**Table 4**, **Supplementary Figure 2**). The number of optical diagnoses made with high confidence also increased (endoscopist alone 92.2% [47/51] vs AI-assisted endoscopists 96.1% [49/51], p=0.500) (**Supplementary Table 2**)."

In the subsequent paragraph, we clarified this further by adding the following sentence:
"Diagnostic performances of the AI-assisted endoscopist were higher compared to both CADx-systems, except for specificity for which CAD EYE performed best."

Discussion – please expand on the limitation that the AI4CRP requires images to be manually captured by a human and the exclusion of some of these images due to being 'out of focus' or 'motion blur'.

We added the following to the limitation section:
"An important limitation was the semi-automated use of AI4CRP. Images had to be manually captured by a research physician, limiting functional use of AI4CRP in clinical practice. A fully automated approach is currently under development. Furthermore, images out of focus or motion blurred imaged were excluded and a new image had to be taken. Although inconvenient, this only hampered the work flow minimally, but could have introduced bias. An image quality indicator alongside the CADx prediction, could be helpful in quantifying and reducing this bias."

Discussion – a comment is made regarding "By comparing a commercially available CADx with an in-house developed CADx, unbiased comparison between the systems and a self-critical system was possible". I do not believe this is entirely true, as from my understanding, the AI4CRP was likely trainined with data from the same site that it was evaluated which can bias the performance to favour AI4CRP (as the CAD-EYE was unlikely trained with data from that site). Please rephrase this sentence to reflect this and add to the limitations that the AI4CRP was only validated at a site from which training data was acquired.

We agree with the reviewer that a bias may exist and may favor AI4CRP performance as it was indeed trained with data from the same hospital in which it was tested. Nevertheless, AI4CRP was also trained with data from five other hospitals. See supplementary table 1 for the complete overview of the training and testing data for AI4CRP. CAD EYE was not trained with data from the testing hospital.

We have rephrased the sentence and have added details about bias in the limitation section:
"Bias could also have occurred since AI4CRP was trained with data from the same hospital in which it was tested in this study, possibly favoring AI4CRP performances, while this is not true for CAD EYE."

Discussion – please rephrase the comment "Both CADx's diagnostic performances approximated the level of the expert endoscopist" to reflect the endoscopist diagnostic performance was non-significantly higher than both CADx-systems

As stated above, it cannot be said that the endoscopists performed non-significantly higher than both CADx-systems.

We rephrased the sentence as follow:
"Diagnostic performances of both CADx-systems were non-significantly inferior compared to the performance of the expert endoscopist, with the exception of specificity, were CAD EYE demonstrated the best performance."

Discussion – the discussion section introduces results for PIVI and SODA but this is not reported in the main manuscript. Unfortunately I do not have access to the supplementary section, please kindly ensure these results are reported there.

We thank the reviewer for the alertness. We have added the explanation on the resect-and-discard and diagnose-and-leave strategies in the introduction and rephrased the section in the discussion:

"Self-critical AI4CRP and CAD EYE reached a NPV of ≥90% for rectosigmoid polyps according to the quality standard for the diagnose-and-leave strategy by the ASGE.[3] Both CADx-systems also met the quality standard of the ESGE for the diagnose-and-leave strategy and self-critical AI4CRP also the ESGE quality standard for the resect-and-discard strategy.[4]"

Discussion - please rephrase the comment "Diagnostic performances of self-critical AI4CRP and CAD EYE approximated the level of the expert endoscopist" to reflect the endoscopist diagnostic performance was non-significantly higher than both CADx-systems
We rephrased the conclusion:
"Diagnostic performances of the AI-assisted endoscopist were higher compared to both CADx-systems."

Editorial office's comments
(1) Science editor
1 Scientific classification: Grade B.
2 Language classification: Grade A.

3 Specific comments:
(1) Please provide the Figures cited in the original manuscript in the form of PPT. All text can be edited, including A,B, arrows, etc. With respect to the reference to the Figure, please verify if it is an original image created for the manuscript, if not, please provide the source of the picture and the proof that the Figure has been authorized by the previous publisher or copyright owner to allow it to be redistributed. All legends are incorrectly formatted and require a general title and explanation for each figure. Such as Figure 1 title. A: ; B: ; C: .
We thank the science editor for the alertness.
We provided the figures in the form of PTT and stated that all images are original and created for this manuscript. We formatted the legends to a general title and a separate explanation for each figure.

(2) Abbreviations other than special types of words such as COVID-19 and SARS-CoV-2 are not allowed in the article title.
We have written out the abbreviation in the title.

(3) Please provide all fund documents.
We uploaded the award decision by the Dutch Cancer Society (KWF) for the funding received for this study.

(4) The "Article Highlights" section is missing. Please add the "Article Highlights" section at the end of the main text (and directly before the References).
We have added the Article Highlights section:

Research background The importance of optical diagnosis, the endoscopic characterization of colorectal polyps, increases. However, correct endoscopic characterization and differentiation between benign and pre-malignant polyps remains difficult even for experienced endoscopists.

Research motivation The ability of modern-day computer-aided diagnosis systems (CADx-systems) to automatically recognize informative patterns in datasets can potentially improve accurate characterization of colorectal polyps and facilitate the implementation of treatment strategies based on optical diagnosis by meeting the set quality standards.

Research objectives Aim of this study was to evaluate the feasibility of the real-time use of the in-house developed CADx-system Artificial Intelligence for ColoRectal Polyps (AI4CRP) for the optical diagnosis of diminutive (≤5 mm) colorectal polyps. Secondary aims were a head-to-head comparison of AI4CRP with CAD EYE™ (Fujifilm, Tokyo, Japan), evaluating the diagnostic performances of self-critical AI4CRP (providing only high confidence diagnoses), the diagnostic performances of an expert endoscopist (endoscopist alone), and the influence of CADx on the optical diagnosis of an expert endoscopist (AI-assisted endoscopist).

Research methods The two CADx-systems (AI4CRP and CAD EYE) were compared head-to-head. Colorectal polyps were characterized as benign or premalignant and histopathology was used as gold standard. AI4CRP provided characterizations accompanied by confidence values, enabling self-critical AI4CRP in which low confidence characterizations were excluded. The AI-assisted endoscopists, optically diagnosing colorectal polyps after reviewing both CADx characterizations.

Research results Real-time use of AI4CRP was deemed feasible in clinical practice. AI4CRP showed a sensitivity of 82.1%, a specificity of 75.0%, a negative predictive value of 56.3%, and a diagnostic accuracy of 80.4%. Self-critical AI4CRP excluded 14 low confidence characterizations, resulted in considerably higher diagnostic performances compared to AI4CRP. CAD EYE had a sensitivity of 74.2%, a specificity of 100.0%, a NPV of 69.2%, and a diagnostic accuracy of 83.7%. Diagnostic performances of the endoscopist alone (before AI) increased non-significantly after reviewing the CADx characterizations of both AI4CRP and CAD EYE (AI-assisted endoscopist). Diagnostic performances of the AI-assisted endoscopist were higher compared to both CADx-systems, except for specificity for which CAD EYE performed best.

Research conclusions Real-time use of AI4CRP was feasible. Objective confidence values provided by a CADx is novel and self-critical AI4CRP showed higher diagnostic performances compared to AI4CRP. Reviewing characterizations by AI4CRP and CAD EYE did not increase the performance of the AI-assisted endoscopist.

Research perspectives Future studies should expand on our findings and further investigate the added value of self-critical CADx-systems.