scientific reports

Check for updates

OPEN

Experimental evidence of effective human–AI collaboration in medical decision-making

Carlo Reverberi[1,2]✉, Tommaso Rigon[3], Aldo Solari[2,3], Cesare Hassan[4,5], Paolo Cherubini[1,2,6], GI Genius CADx Study Group* & Andrea Cherubini[2,7]✉

Artificial Intelligence (AI) systems are precious support for decision-making, with many applications also in the medical domain. The interaction between MDs and AI enjoys a renewed interest following the increased possibilities of deep learning devices. However, we still have limited evidence-based knowledge of the context, design, and psychological mechanisms that craft an optimal human–AI collaboration. In this multicentric study, 21 endoscopists reviewed 504 videos of lesions prospectively acquired from real colonoscopies. They were asked to provide an optical diagnosis with and without the assistance of an AI support system. Endoscopists were influenced by AI (or = 3.05), but not erratically: they followed the AI advice more when it was correct (or = 3.48) than incorrect (or = 1.85). Endoscopists achieved this outcome through a weighted integration of their and the AI opinions, considering the case-by-case estimations of the two reliabilities. This Bayesian-like rational behavior allowed the human–AI hybrid team to outperform both agents taken alone. We discuss the features of the human–AI interaction that determined this favorable outcome.

Artificial Intelligence systems are increasingly recognized as a precious tool for improving medical-decision making[1]. AI may support Medical Doctors (MDs) in multiple domains (with applications in dermatology, ophthalmology, cardiology, gastroenterology, and mental health, among others) while typically MDs keep the final decision. Such complementarity advises a collaboration between human and artificial minds, fostering a "hybrid intelligence" that could deliver outcomes superior to those reached by each mind alone[2,3]. Pivotal to fulfilling this promise is improving the interaction between humans and machines to build up an effective team avoiding pitfalls such as:

1. *over-reliance*: MDs adhere to whatever opinion is offered by the AI, ignoring their independent evaluation. This attitude throws away all the information embedded in the MD's own opinion and could endanger the accuracy of the final diagnosis. Previous studies on human trust toward AI decision-support systems alerted us of the possibility of an extreme form of over-reliance, termed "automation bias" for AI's false alarms and "automation complacency" for AI's false reassurances. In both cases, the humans uncritically adhere to the machine's output, ignoring their independent evaluation[4–6].
2. *under-reliance*: MDs display limited trust in the machine and mostly ignore its suggestions, even when informative. If this attitude were dominant, AI would prove useless to any practical means. The extreme form of under-reliance is termed "algorithm aversion"[4,7,8]: the human does not trust the machine and completely ignores its suggestions.
3. *opacity* of judgments' reliability: in this case, even if MDs have an appropriate level of trust towards AI, they cannot tell whether AI opinions are more or less reliable than their own. Opacity may prevent MD from reaching an optimal use of the information provided by the AI. Previous studies have addressed this topic by trying to convey the AI's internal motives of AI decisions, but with mixed results[6,9].

[1]Department of Psychology, University of Milano-Bicocca, 20126 Milan, Italy. [2]Milan Center for Neuroscience, University of Milano-Bicocca, 20126 Milan, Italy. [3]Department of Economics, Management and Statistics, University of Milano-Bicocca, 20126 Milan, Italy. [4]Department of Biomedical Sciences, Humanitas University, 20072 Pieve Emanuele, Italy. [5]Endoscopy Unit, Humanitas Clinical and Research Center IRCCS, Rozzano, Italy. [6]Department of Neural and Behavioral Sciences, University of Pavia, Pavia, Italy. [7]Artificial Intelligence Group, Cosmo AI/Linkverse, Lainate, 20045 Milan, Italy. *A list of authors and their affiliations appears at the end of the paper. ✉email: carlo.reverberi@unimib.it; acherubini@cosmopharma.com

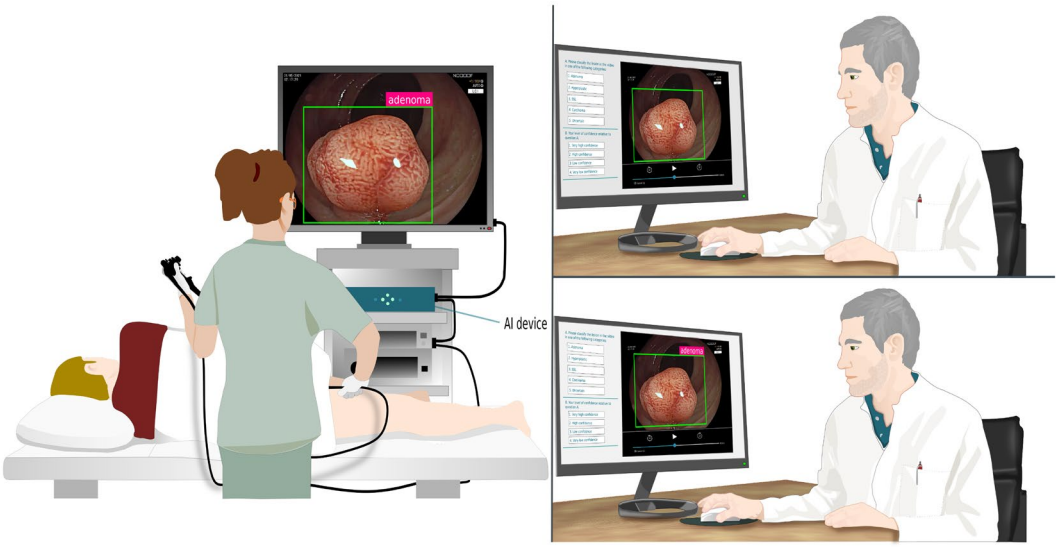

Figure 1. Left panel: The stimuli used in the experiment were prospectively collected in a real-world clinical setting using an AI medical device supporting MDs for lesion detection (CADe) and categorization (CADx) as adenomatous or non-adenomatous[24]. Right panel: An international group of endoscopists were asked to optically diagnose the same set of lesions, presented as short video clips, in two experimental sessions. In the first session (top-right panel) the AI only highlights the target lesion, while in the second session (bottom-right panel) AI also dynamically offers an optical diagnosis. For more details on the AI device see Appendix A.1.2.

The success and the potential drawbacks of human–AI collaboration are under active scrutiny in the clinical domain and beyond. A renewed interest on this topic followed the new possibilities granted by deep learning tools, thus generating several related lines of research (e.g. "augmented intelligence", "hybrid intelligence", "human–AI collaboration", "human–AI teaming")[2,3,10–14]. Notwithstanding a clear agenda on the issues that we should explore further, we still have limited evidence-based knowledge of the context, design, and psychological mechanisms that would craft an optimal human–AI team[15–17]. Many experimental studies and reviews measured the performance of AI-based medical devices, or the improvement of the MDs diagnostic accuracy when supported by AI (e.g.,[18–22]). One review focused specifically on radiologists trust in AI's recommendations[23]. To our knowledge, no previous experimental study addressed the inner dynamics of AI-supported MDs' beliefs revision. AI assisted colonoscopy provides a privileged case study on team-working between humans and machines. The colonoscopic procedure naturally emphasizes the complementary roles of the endoscopist and the AI (Fig. 1). On one side, the endoscopist is fundamental for navigating the probe and selecting the information. On the other hand, the limited time available combined with the intense multitasking and the multidimensional nature of lesion classification imply the potential benefit of additional help.

We study the interaction of endoscopists and a last-generation decision support system[24] during the optical examination of colorectal lesions. We model the endoscopist diagnosis as a psychological categorization process[25,26]. Incoming visual information is compared to previously stored knowledge about a finite set of possible diagnoses in a Bayesian-like procedure that revises the endoscopist's confidence toward those diagnoses. Accrued information can make one diagnosis dominant over its alternatives. We aim to understand how the availability of the AI advice influences the opinion of the endoscopist (i.e., whether an "effective hybrid team" is formed) and whether individual features like the endoscopist expertise modulate the AI influence. In qualitative Bayesian terms, the AI output is a further piece of information that should be integrated to revise the endoscopist's diagnosis, not differently as it would happen by considering an "informed opinion" volunteered by a human colleague with a slightly different expertise profile. If such well-calibrated interaction is achieved, endoscopists' accuracy with AI should improve, notwithstanding their baseline level of accuracy without AI, because they would have available one further piece of information (Fig. 2).

Expert and novice endoscopists were asked to optically diagnose colorectal lesions. The same set of lesions ($n = 504$) was presented, as short video clips, in two experimental sessions. In the first session the AI (Fig. 1) only highlights the target lesion, while in the second session AI also dynamically offers an optical diagnosis. We had four leading experimental questions, i.e., whether endoscopists are influenced by the AI opinion, and in case, whether this leads to an improvement of diagnostic accuracy; whether the endoscopists could selectively follow the AI when it is correct and conversely reject AI's opinion when it is incorrect. Overall, we hypothesized that endoscopists consider AI's opinion to improve their diagnostic performance and can discriminate correct from potentially wrong AI's opinions. Further planned measures and analyses aimed at clarifying the reasons underlying the endoscopists' behavior and the interaction between humans and AI. Namely, we hypothesized that endoscopists have a reliable insight on the correctness likelihood of their own and AI's opinions and that they use such insight to weight the human and AI judgment. Finally, a larger increase in accuracy should be observed for the non-expert endoscopists. Our predictions were pre-registered before the data gathering, together with the study plan, the statistical models, and the analyses (preregistration is available at https://osf.io/y9at5).

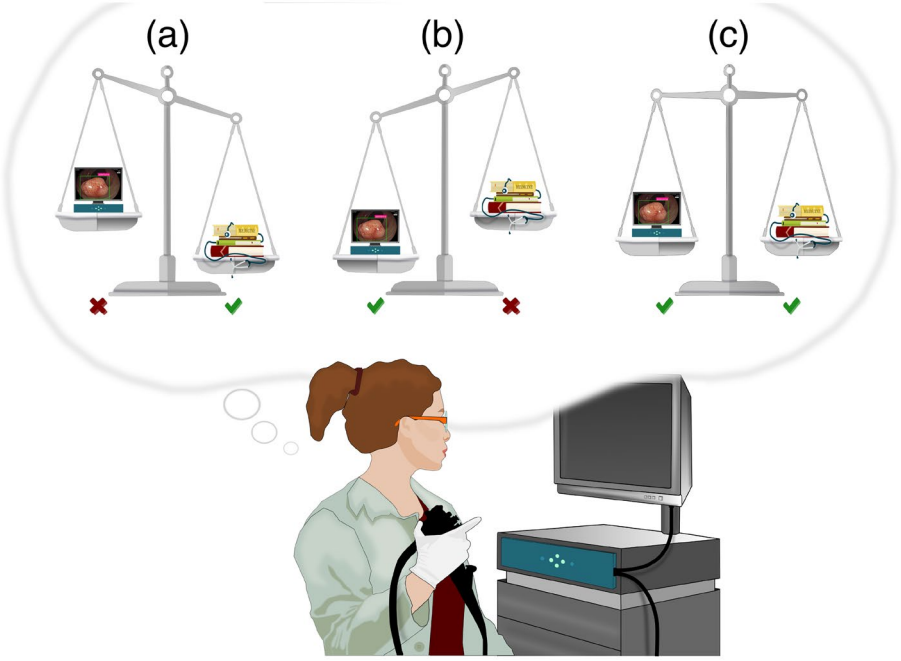

Figure 2. MD-AI team. An endoscopist subject to under-reliance discounts the added information given by the AI (**a**). An endoscopist subject to over-reliance supinely accepts the AI suggestion (**b**). The optimal use of AI should rest on an in-between, well-calibrated approach where the endoscopist uses the AI opinion for coherently revising their confidence in their initial evaluation. In this way, the medical decision-making process would benefit from a collaboration between the two intelligences (**c**).

Our study is an original contribution to the literature in several ways. First, we investigated whether and how the output of an AI real-time classifier influences the decision of a MD in a within-subject design that compares decisions of the same MD, with and without AI support, on a prospectively acquired dataset of colonoscopy videos. Previous studies on CADx systems for colonoscopy focused on assessing the performances of AI against the accuracy of physicians[19,20,27–30], or against the criteria established by gastroenterology societies for implementing the technology in a clinical setting[31–34].

Second, we developed a new, rigorous statistical model for measuring MDs' belief revisions in experimental settings by framing the MDs' diagnostic updates following AI advice as a Bayesian-like belief-revision process. The model separates "efficacy" (whether the MD aligns her belief to the AI when the latter is correct) and "safety" (whether the MD stays with her previous belief when the AI is wrong). Thus the model transparently assesses the positive and negative impact of AI opinions on MD decisions and their ability to avoid over/under reliance.

Third, and importantly, we explored the psychological processes underlying the emergence of an effective hybrid team, even when humans need to interact with a non-transparent AI device (i.e., a device not conveying the motives of its decisions). For that aim, we collected the critical parameters that should contribute to a final MDs decision: the MDs opinion, their confidence, their interpretation of the AI output, and the perceived reliability of each AI advice. These parameters were ignored in previous studies on AI medical devices.

Methods

Study design and participants. This is an experimental study with a mixed 2 × 2 design, in which the within-subjects factor is the treatment (no AI vs. AI), and the between-subjects factor is the endoscopists level of expertise (experts vs. non-experts). Twenty-one endoscopists took part in the experiment, 10 of them experts (at least 5 years of colonoscopy experience and experience in optical biopsy with virtual chromoendoscopy) and 11 non-expert (less than 500 colonoscopies performed). The acquired sample size ensures adequate statistical power for testing our main experimental endpoints (the power analysis is in Appendix A.5). The participants were from Austria, Israel, Japan, Portugal, and Spain (see the consortium members at the end of the paper). Both the task and task instructions were in English. All methods were carried out, and results were reported following STROBE guidelines and regulations. The study was approved by the local Institutional Review Board (Comitato Etico Lazio 1, prot. 611/CE Lazio1) and conducted following the Declaration of Helsinki. All participants provided written informed consent.

Procedure and data collection. The experiment is divided into two sessions. In *Session 1* (S1), endoscopists diagnose the lesions without AI advice. Endoscopists watched on an online dedicated platform 504 videos of real colonoscopies, each presenting one lesion. Endoscopists examined lesions at their own pace. Their task was, first, to categorize each lesion in five forced-choice options: "Adenoma", "Hyperplastic", "SSL",

Variable name	Description
Histologic evaluation	The ground truth of each lesion. Its possible values were: "Adenoma", "Non-Adenoma"
Human judgment, S1 and S2	The optical diagnosis of the lesion by an endoscopist in each session, mapped as mentioned above. It takes the values: "Adenoma", 'Non-adenoma", and "Uncertain"
Human confidence, S1 and S2	The confidence of the previous judgment expressed by the endoscopist. It takes the values: "Very high", "High", "Low", "Very low", "Uncertain". We classified the confidence as "Uncertain" whenever the associated lesion evaluation was "Uncertain"
Algorithmic AI diagnosis	The diagnosis about a given lesion provided by the AI output in S2, as interpreted by an automatic algorithm. It takes the values: "Adenoma", "Non-Adenoma", and "Undetermined"
Perceived AI diagnosis	The endoscopists' interpretation of the AI dynamic output in S2. It takes the following values: "Adenoma", "Non-Adenoma", "Uncertain", and "I am not sure/I did not notice the output of the AI"
Evaluation of AI confidence	The endoscopists' appreciation of the level of reliability of the AI diagnosis in S2. It takes the following values: "Very high confidence", "High confidence", "Low confidence", and "Very low confidence"
Transformed variable name	
Human correct diagnosis, S1 and S2	A binary variable that indicates whether each lesion was correctly diagnosed by each endoscopist, in each session
Accuracy$_s$ of the perceived AI diagnosis	A binary variable indicating whether each lesion is correctly diagnosed by the AI. We considered the AI diagnosis perceived by the endoscopist. "Uncertain" and "not sure/not notice" were excluded from the computation
Accuracy$_u$ of the perceived AI diagnosis	A binary variable indicating whether each lesion is correctly diagnosed by the AI. We considered the AI diagnosis judged by the endoscopist. Classifications of the AI output as "Uncertain" and "not sure/not notice" were conservatively considered errors
Confidence score, S1 and S2	A discrete numerical variable ranging 1 to 9 that measures the belief of each endoscopist about each lesion in each session. The score of 9 indicates a strong belief that the lesion is an adenoma. At the other extreme, the score of 1 denotes a strong belief that the lesion is not an adenoma. The score of 5 indicates "Uncertain" diagnoses

Table 1. Measured and transformed variables for each of the 21 subjects and 504 lesions.

"Carcinoma", "Uncertain". Their choices were later mapped to the 3-fold output "Adenoma" (corresponding to the choices "Adenoma" or "Carcinoma"), "Non-Adenoma" (for "Hyperplastic" or "SSL" choices), or "Uncertain". The second task was to describe their confidence in their decision as: "Very high", "High", "Low", or "Very low". We recorded the time elapsed from the beginning of the video to the first decision. No feedback was provided. In *Session 2* (S2), endoscopists' decision was supported by AI: the endoscopists saw the same videos as in S1, but with the AI's dynamic advice for each lesion: the AI's optical diagnosis for a specific lesion could vary between frames, displaying one of the four possible values "adenoma", "non-adenoma", "no-prediction" or "analyzing" (see Appendix A.1.2). The task and lesions were the same as in S1. However, besides the two questions identical to S1, in S2 the endoscopists had to also report the perceived AI's overall opinion about the lesion, using 3 forced-choice response options: "Adenoma", "Non-Adenoma", and "Uncertain"; and the perceived AI's level of confidence about the lesion, using 4 forced-choice response options: "Very high", "High", "Low", and "Very low". The 504 videos in each session were divided into 6 batches of 84 videos each, with a predefined sequence of administration of the batches. For each batch, we prepared different pre-randomized orders of presentation of the lesions. Each participant was preassigned to a different order. At least two weeks passed from the conclusion of the evaluation of one batch in S1 and the evaluation of the same batch in S2 to avoid memory effects. All stimuli were acquired in full length and with no AI overlay in the *CHANGE* clinical study[34]. For generating S1 clips, we used GI Genius v3.0 in CADe modality to dynamically add a green box around the suspect lesions automatically detected by the device (see Appendix A.1.2 for a detailed description of the AI system). For S2 we used GI Genius v3.0 in CADe+CADx modality[24], dynamically overlaying a green box around suspected lesions and the optical diagnosis computed by the AI. For evaluating human and AI performance, we considered the histopathological diagnosis of each lesion as ground truth. A more detailed description of stimuli and procedure is in the Appendix A. Three example stimuli are available online; for a description of the video clips, see Appendix A.1.1. The recorded and transformed variables are described in Table 1.

Statistical analyses. For comparing the probability of prediction-relevant events during S1 with the probability of the same events during S2, for each endoscopist, we defined and computed four odds ratios amenable to be analyzed by a logistic regression statistical model (see Appendix A.4). The four main experimental endpoints and the related odds ratios are:

1. AI *influence* on endoscopists' decision (ω_I): the ratio between the odds that the endoscopists' diagnoses were the same as AI's diagnoses in S2, and the odds that the endoscopists' diagnoses in S1 were the same as those given by AI on the same lesions in S2, irrespective to accuracy. Values greater than 1 mean that in S2 the endoscopists' diagnoses converged on AI diagnoses. The opinion of the endoscopist should be influenced by the response of the AI, i.e., we hypothesize $\omega_I > 1$.
2. AI *effect on diagnostic accuracy* (ω_A): computed as a ratio between the odds that the endoscopists' diagnoses were correct in S2, and the odds that the endoscopists' diagnoses were correct in S1. Values greater than 1

Endpoint	OR	Estimate
1. Influence of the AI	ω_I	3.05 [2.76, 3.39]
2. Diagnostic accuracy	ω_A	1.39 [1.28, 1.51]
3. Effectiveness	ω_E	3.48 [3.07, 3.98]
4. Safety	ω_S	0.54 [0.48, 0.62]

Table 2. Odds-ratios (OR) for each of the main endpoints. We report in brackets the 95% confidence intervals for the odds ratios.

mean that AI's assistance is associated with an increased probability of a correct human diagnosis. Accuracy should improve with AI, thus $\omega_A > 1$.

3. *Effectiveness* (ω_E): same as ω_A, but using odds estimated only on the subset of lesions where AI returned a correct diagnosis. Values greater than 1 mean that when AI is correct, AI's assistance increases the probability of a correct human diagnosis. The endoscopists should rightfully accept the AI opinion when this is correct, thus $\omega_E > 1$.
4. *Safety* (ω_S): same as ω_A, but using odds estimated only on the subset of lesions where AI returned a wrong diagnosis. Values less than 1 mean that when AI is incorrect, AI's assistance deteriorates the endoscopists' diagnostic performance. The endoscopists should be able to disengage from a wrong opinion of the AI so that their `S1` performance is not remarkably deteriorated when AI's offers a wrong suggestion. Thus, we hypothesized $\omega_S > 0.3$.

The above four sets of odds ratios $\omega_I, \omega_A, \omega_E, \omega_S$ are obtained as the result of logistic regression models that account for lesion- and endoscopist- specific random effects. The parameters and the corresponding odds ratios of the logistic regressions were estimated via (integrated) maximum likelihood and are based on the `lme4`[35]. Related inferential quantities (e.g., confidence intervals) were also computed using the `lme4` R package. The complete mathematical definitions of the transformed variables, the odds ratios, the relative risks, the statistical models used, and the details of the inferential tests run are available online in Appendix A.4, A.3, and A.6.

Results

Our main, pre-registered expectations were fully supported by results (Table 2, Fig. 3, see section "Statistical analyses" for details on the measures used). Endoscopists were influenced by the AI opinion ($\omega_I = 3.05$). On average, for every three lesions over which endoscopists disagreed with AI in `S1`, only one remained in `S2`. Considering AI opinion is beneficial to the diagnostic performance of the endoscopists: every five lesions correctly evaluated without AI, seven were correctly evaluated with AI. Importantly, endoscopists could discriminate the good from the bad AI advice. When the AI was correct, the endoscopists followed its advice more ($\omega_E = 3.48$) than when the AI was incorrect ($1/\omega_S = 1.85$). In other words, endoscopists could not fully escape from the negative consequences of a wrong AI advice (i.e., ω_S is less than 1), but this was more than compensated by their stronger tendency to accept a correct AI advice.

Additional analyses: explaining endoscopist behavior. Overall, additional analyses aimed at understanding the reasons underlying the observed endoscopists' behavior. We group additional results into two sets: endoscopists' assessment of the main task parameters, AI and expertise effects. Further results (e.g., time to decision, individual-level performance) and alternative analysis approaches (e.g., ROC curves) are reported in Appendix B.

Endoscopists' assessment of the main task parameters. We hypothesized that the endoscopists are aware of the changing soundness of their judgments and are naively able to interpret the AI output and assess its reliability. To test these hypotheses, we considered three measures: the confidence of the endoscopists over their diagnosis, the endoscopists' interpretation of the AI output, and its confidence. We found that endoscopists' confidence in an optical diagnosis is strongly predictive of its accuracy in both sessions: the higher the confidence, the higher the accuracy (Table 3). The endoscopists' interpretation of the AI output is overall consistent with an algorithmic assessment of the AI output (81% of agreement between endoscopists and AI, 94% agreement when judgments "uncertain" are excluded). It is consistent across endoscopists (77% of average agreement among all the possible pairs of endoscopists, 94% when "uncertain" are excluded). Finally, and importantly, the endoscopists' estimates of the AI reliability are predictive of AI accuracy (Table 3).

Having shown that endoscopists can generate meaningful estimates of critical decision parameters, we asked whether they used them to optimally integrate their opinion with that of AI. An intuitive Bayesian decision-maker would weigh each opinion over its estimated reliability. We thus asked whether the influence of AI on endoscopists' decisions would change depending on their own and AI confidence. We found that this is indeed the case: when endoscopist confidence (as estimated in `S1`) is high, or AI confidence is low, the endoscopists tend to stick with their decision, i.e., they do not change their mind in case of disagreement with AI (Table 4). The other way around in case of low endoscopist confidence or high AI confidence. Similar results were obtained in a related analysis on confidence scores (see Appendix B.3).

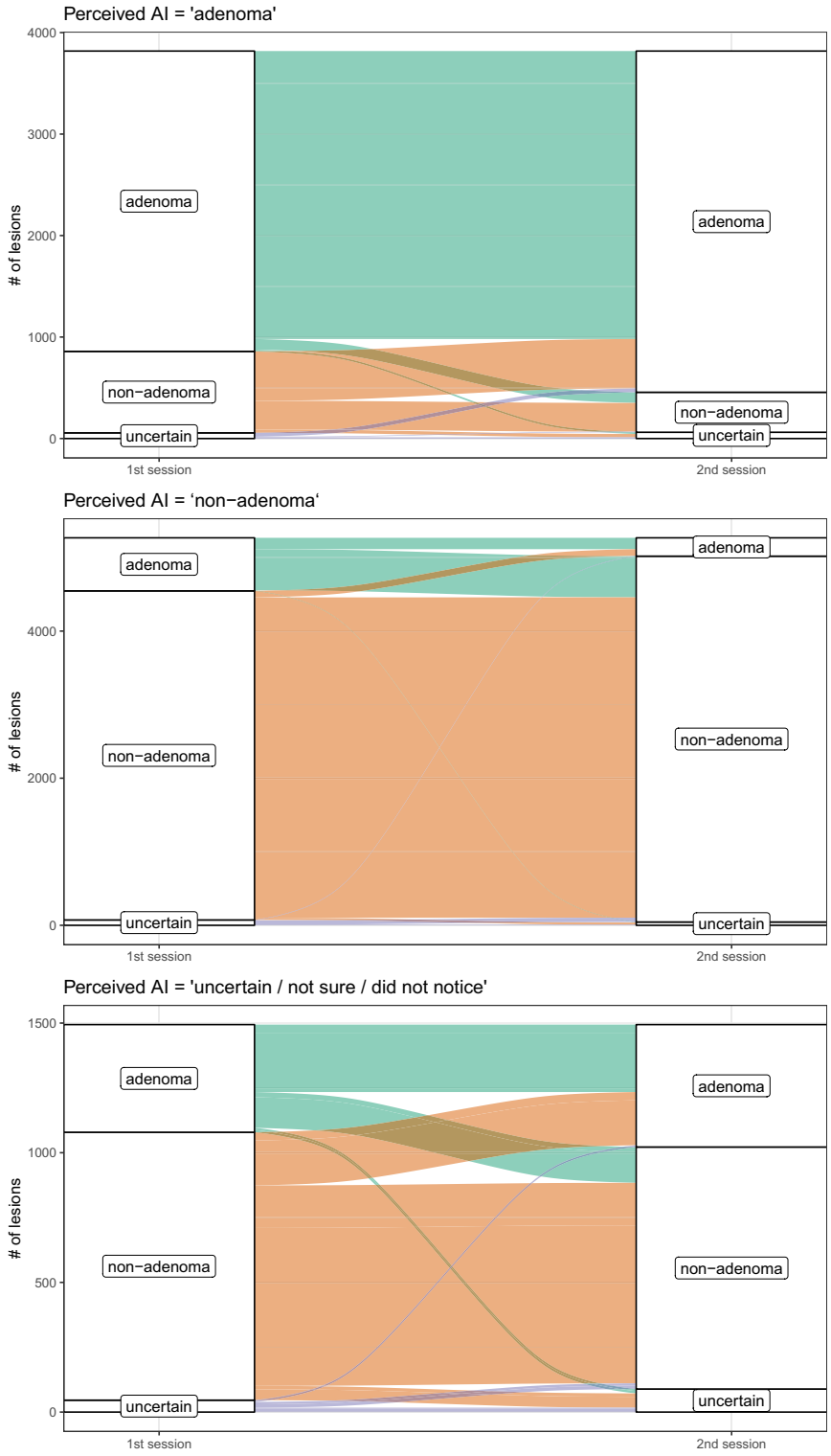

Figure 3. Influence of the AI: alluvial diagrams representing changes in endoscopist's opinion between the two sessions as a function of perceived AI response.

Effects of expertise. Endoscopists' expertise modulated the variables considered for our main analyses. AI affected non-experts more than experts, and it had a larger impact on accuracy, possibly because experts' had less to gain from AI's added information since their accuracy was close to the AI accuracy. Furthermore, experts were less able than non-experts to discriminate between good and bad AI advice: both efficacy and safety are higher in non-experts than in experts (Table 5). The lower safety of experts seems at first counter-intuitive. However, it may be understood as a stronger preference of experts to avoid false-negative errors at the cost of

Confidence	Very low	Low	High	Very high	Overall
S1 accuracy	0.644 (236)	0.685 (2665)	0.806 (5263)	0.853 (2241)	0.768 (10,584)
S2 accuracy	0.543 (184)	0.679 (1859)	0.839 (5235)	0.882 (3094)	0.802 (10,584)
AI $accuracy_s$	0.667 (216)	0.718 (1456)	0.863 (4608)	0.909 (2807)	0.849 (9086)

Table 3. Proportions and sample sizes of the correct human diagnosis in S1, S2 and the AI perceived correct diagnosis, against different human confidence levels and the AI perceived confidence levels, respectively. Following the standard in the field, $accuracy_s$ does not consider wrong lesions where AI opinion was perceived as "uncertain" or was "not noticed". Evaluations of the confidence of the AI were asked only when the opinion was "Adenoma" or "Non-Adenoma".

	Very low	Low	High	Very high
Human conf.	0.703 (64)	0.738 (577)	0.668 (689)	0.598 (194)
AI conf.	0.278 (97)	0.438 (457)	0.827 (684)	0.888 (286)

Table 4. Change in *agreement* between endoscopists' and AI, measured as the amount of times each endoscopist changes its opinion and follows AI's suggestion. We report proportions and sample sizes of the change in *agreement* for different human confidence levels (S1) and the AI perceived confidence levels, respectively.

Endpoint	OR	Expert	Non-expert
1. Influence of the AI	ω_I	2.88 [2.48, 3.34]	3.20 [2.80, 3.65]
2. Diagnostic accuracy	ω_A	1.15 [1.01, 1.30]	1.61 [1.44, 1.79]
3. Effectiveness	ω_E	3.22 [2.64, 3.93]	3.65 [3.11, 4.28]
4. Safety	ω_S	0.45 [0.37, 0.54]	0.63 [0.54, 0.75]

Table 5. Odds-ratios (OR) for each endpoint, estimated separately for experts and non experts. We report in brackets the 95% confidence intervals.

increasing false positives: experts accept more than non-experts an incorrect "adenoma" advice from AI. This is arguably a clinically prudent approach in colonoscopies. The interpretation is supported by the observation that in S2 non-experts increased both their specificity and sensitivity, whereas experts increased only their sensitivity (see Appendix B.5).

The average confidence of experts was lower than those of non-experts, in all sessions, both towards their own judgments and towards AI output. However, no differences were present in the relative trust towards AI in the two groups: both groups had a slightly higher average confidence towards AI as compared to themselves (see Fig. B.11 in Appendix). This observation implies that confidence cannot explain the different attitudes of experts towards AI. More importantly, the confidence of both non-experts and experts was predictive of real accuracy (own or AI). This finding means that both expertise groups can appropriately use confidence to inform their final decision.

AI performance. Depending on how one interprets the rich, dynamic output of the AI, the AI accuracy would change. To provide a fuller picture, we report the AI accuracy in multiple ways. First, we considered the human interpretation of the AI output: the AI standard accuracy ($accuracy_{ss}$), which excludes "uncertain" or "not sure/not notice" outputs from consideration, is 84.9%. A more conservative accuracy that includes also "uncertain" and "not sure/not notice" outputs as errors ($accuracy_u$) is 72.9%. On average, we observed 71 AI outputs classified by the endoscopists as "uncertain" or "not sure/not notice" out of 504. The AI accuracy based on human interpretations varies across individuals: see Appendix B.2 for details. Second, when an automated algorithm interprets the output, the AI accuracy is 84.5% when the label "uncertain" is excluded, while it is 79.3% when "uncertain" is considered an error.

Discussion

AI systems are increasingly considered for supporting and improving the medical decision process. However, in many scenarios AI cannot (or should not) substitute the human professional[1,36]. Conversely, what is envisaged is teaming humans together with artificial intelligence to exploit the advantages of hybrid intelligence[2,3]. Would this union be able to capitalize on the respective strengths? Which are the potential factors enabling an optimal interaction? Optical diagnosis during colonoscopy represents a telling case study for answering these questions.

Endoscopists' optical diagnosis is a psychological categorization process. Endoscopists use incoming visual information to generate possible alternative diagnoses and revise their confidence toward each. The availability

of the AI advice is one more piece of information that endoscopists may actively use in this Bayesian-like revision process. We showed that endoscopists consider and are substantially influenced by AI opinion. Importantly, endoscopists can separate the good from the bad AI advice, accepting selectively more the former than the latter, as shown by a higher efficacy than safety index. This ability, combined with the relatively high accuracy of the AI classification (∼ 85%), granted a beneficial effect on the overall diagnostic performance: the "hybrid human–AI teams" had, on average, better accuracy than the endoscopists alone.

How did the endoscopists select and follow the best AI advice? The successful extraction of two critical task parameters was likely at the core of this ability: endoscopists could intuitively but reliably predict for each lesion both their accuracy (not obvious[37,38]) and the accuracy of the AI (not obvious[39]). Furthermore and importantly, these prediction estimates affected endoscopists' decisions so that they switched their diagnosis towards the AI opinion more when their confidence was low and AI perceived confidence was high. Vice-versa, endoscopists stuck with their diagnosis when their confidence was high and AI perceived confidence was low. In this way, belief revision turned out to be a sound practice in Bayesian terms, resulting in overall increased accuracy of the human–AI hybrid team.

A key to the success of hybrid teams is to calibrate human trust in AI for each specific decision. Knowing when to trust or distrust the AI allows the human expert to apply its knowledge appropriately, improving decision outcomes in cases where the AI is likely to perform poorly[4,40]. Three pitfalls undermine the beneficial effects of human–AI interaction. The first two, over-reliance or under-reliance on AI, regard a general attitude towards support systems, which is wrong when decoupled from considerations on the relative informativeness of the AI[39,41,42]. The third pitfall is more subtle and pervasive: opaque reliability of AI or human judgments, i.e., the MD might not know how much s/he can trust her own, or the AI's, judgment in each specific medical problem. If the case-by-case reliability of judgments is unknown or miscalibrated, a correct Bayesian-like belief revision could be severely hindered. Our study shows that none of these potentially dangerous patterns occurred. The endoscopists were able to build a correct mental model of the AI's error boundaries[43]. They did so by capitalizing on explicit general warnings of AI accuracy but, more importantly, on cues in the AI output, spontaneously interpreted as informative on the AI's confidence in its diagnosis: the persistence of the same AI diagnosis, and the rate of no-prediction AI output (Appendix B.4). MD seemed to evaluate AI confidence as they would have evaluated a colleague's by perceiving his/her hesitation[44].

Our results could generalize to different medical settings. The given interpretation of the success of our hybrid teams stresses one enabling ingredient on the AI side, namely to provide the MDs with an intuitive - yet valid - clue to AI reliability. This finding should alert decision support systems developers: AI's perceived confidence that was unrelated to its accuracy and error boundaries[45] might fool into error the human side of the decision team[46]. On the other hand, the absence of algorithmic transparency in the AI device considered in this study did not have a disruptive effect on MD performance, arguably because MDs could infer AI reliability from other indirect cues. Thus, even though algorithmic transparency has been sometimes advocated as pivotal for promoting an effective interaction with AI[6,9,36], we suggest that easy access to case-by-case reliability may be a sufficient, or even more important, factor.

As expected, expert endoscopists showed a better performance in optical diagnosis overall standard descriptive parameters: accuracy, sensitivity, and specificity (Appendix B.2). In this context, however, the most important - and reassuring - finding was the ability to interact intuitively with AI shared between *both* expert and non-expert. The main results on influence, safety, and efficacy held for both subgroups. The benefits on accuracy were stronger for non-experts (also given their lower baseline), making their performance with AI assistance similar to that of experts without assistance. These observations also open up the interesting possibility of using AI systems for juniors' training.

In high-stakes scenarios, such as in the medical domain, full automation of decision-making is often impossible or undesirable. This is not only for ethical or regulatory issues but also because human experts can rely on their domain knowledge, complementary to the AI's. In hybrid decision making, the individual strengths of the human and the AI come together to optimize the joint decision outcome. In the present case study, the use of AI proved effective and safe. Effective, because MDs adhered to AI's opinions mostly when the latter were correct. Safe, because MD's adherence to AI opinions was relatively low when the latter were incorrect. When these enabling conditions are met, hybrid decision-making is an effective and appropriate diagnostic approach.

From these conclusions, we can distillate two leading suggestions. To physicians: treat AI opinion as you would treat advice from a human colleague with a slightly different expertise profile: weigh advice based on the relative historical performance between you and the colleague (i.e., how good AI has proven to be in general compared to you), but also on the colleague confidence/hesitation on the specific case. To device manufacturers: make the case-by-case confidence of the device output intuitively readable to the user.

Data availability

To improve transparency and reproducibility of our work, raw data with de-identified participants are publicly available at OSF https://osf.io/57smj.

Received: 10 March 2022; Accepted: 18 August 2022
Published online: 02 September 2022

References

1. Topol, E. J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
2. Dellermann, D. *et al.* The future of human–AI collaboration: A taxonomy of design knowledge for hybrid intelligence systems. arXiv:2105.03354 (2021).

3. Akata, Z. *et al.* A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* **53**, 18–28 (2020).
4. Zhang, Y., Liao, Q. V. & Bellamy, R. K. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305 (2020).
5. Wickens, C. D., Clegg, B. A., Vieane, A. Z. & Sebok, A. L. Complacency and automation bias in the use of imperfect automation. *Hum. Factors* **57**, 728–739 (2015).
6. Gretton, C. Trust and transparency in machine learning-based clinical decision support. In *Human and Machine Learning* 279–292 (Springer, 2018).
7. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**, 114 (2015).
8. Dietvorst, B. J., Simmons, J. P. & Massey, C. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Manag. Sci.* **64**, 1155–1170 (2018).
9. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
10. Park, S. Y. *et al.* Identifying Challenges and Opportunities in Human-AI Collaboration in Healthcare. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, 506–510 (2019).
11. Wang, D. *et al.* Designing AI to work WITH or FOR people? In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–5 (2021).
12. Bansal, G., Nushi, B., Kamar, E., Horvitz, E. & Weld, D. S. Is the most accurate AI the best teammate? Optimizing AI for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 11405–11414 (2021).
13. Wang, D. *et al.* From human–human collaboration to human–AI collaboration: Designing AI systems that can work together with people. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–6 (2020).
14. Bazoukis, G. *et al.* The inclusion of augmented intelligence in medicine: A framework for successful implementation. *Cell Rep. Med.* **3**, 1–8 (2022).
15. Cabitza, F., Campagner, A. & Simone, C. The need to move away from agential-AI: Empirical investigations, useful concepts and open issues. *Int. J. Hum. Comput. Stud.* **155**, 1–11 (2021).
16. Okamura, K. & Yamada, S. Adaptive trust calibration for human–AI collaboration. *PLoS One* **15** (2020).
17. Gu, H., Huang, J., Hung, L. & Chen, X. A. Lessons learned from designing an AI-enabled diagnosis tool for pathologists. In *Proceedings of the ACM on Human–Computer Interaction*, Vol. 5, 1–25 (2021).
18. Aziz, M., Fatima, R., Dong, C., Lee-Smith, W. & Nawras, A. The impact of deep convolutional neural network-based artificial intelligence on colonoscopy outcomes: A systematic review with meta-analysis. *J. Gastroenterol. Hepatol.* **35**, 1676–1683 (2020).
19. Kudo, S.-E. *et al.* Artificial intelligence and colonoscopy: Current status and future perspectives. *Dig. Endosc.* **31**, 363–371 (2019).
20. Larsen, S. L. V. & Mori, Y. Artificial intelligence in colonoscopy: A review on the current status. *DEN Open* **2** (2022).
21. Taghiakbari, M., Mori, Y. & von Renteln, D. Artificial intelligence-assisted colonoscopy: A review of current state of practice and research. *World J. Gastroenterol.* **27**, 8103 (2021).
22. Nagendran, M. *et al.* Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *Br. Med. J.* **368** (2020).
23. Jorritsma, W., Cnossen, F. & van Ooijen, P. M. Improving the radiologist–CAD interaction: Designing for appropriate trust. *Clin. Radiol.* **70**, 115–122 (2015).
24. Biffi, C. *et al.* A novel AI device for real-time optical characterization of colorectal polyps. *npj Digit. Med.* **5**, 1–8 (2022).
25. Griffiths, T. L., Kemp, C. & Tenenbaum, J. B. Bayesian models of cognition. In *The Cambridge Handbook of Computational Psychology* (Cambridge University Press, 2008).
26. Anderson, J. R. The adaptive nature of human categorization. *Psychol. Rev.* **98**, 409 (1991).
27. Mori, Y. *et al.* Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: A prospective study. *Ann. Intern. Med.* **169**, 357–366 (2018).
28. Byrne, M. F. *et al.* Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut* **68**, 94–100 (2019).
29. Xu, Y. *et al.* Comparison of diagnostic performance between convolutional neural networks and human endoscopists for diagnosis of colorectal polyp: A systematic review and meta-analysis. *PLoS One* **16**, e0246892. https://doi.org/10.1371/journal.pone.0246892 (2021).
30. Kudo, S.-E. *et al.* Artificial intelligence and computer-aided diagnosis for colonoscopy: where do we stand now? *Transl. Gastroenterol. Hepatol.* **6** (2021).
31. ASGE Technology Committee *et al.* ASGE Technology Committee systematic review and meta-analysis assessing the ASGE PIVI thresholds for adopting real-time endoscopic assessment of the histology of diminutive colorectal polyps. *Gastrointest. Endosc.* **81** (2015).
32. Berzin, T. M. *et al.* Position statement on priorities for artificial intelligence in GI endoscopy: A report by the ASGE Task Force. *Gastrointest. Endosc.* **92**, 951–959. https://doi.org/10.1016/j.gie.2020.06.035 (2020).
33. Barua, I. *et al.* Real-time artificial intelligence-based optical diagnosis of neoplastic polyps during colonoscopy. *NEJM Evid.* **1** (2022).
34. Hassan, C., Balsamo, G., Lorenzetti, R., Zullo, A. & Antonelli, G. Artificial intelligence allows leaving-in-situ colorectal polyps. *Clin. Gastroenterol. Hepatol.*https://doi.org/10.1016/j.cgh.2022.04.045 *(2022).*
35. Bates, D., Martin, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–47 (2015).
36. WHO. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance* (World Health Organization, 2021).
37. Naguib, M. *et al.* Anesthesiologists' overconfidence in their perceived knowledge of neuromuscular monitoring and its relevance to all aspects of medical practice: An international survey. *Anesth. Analg.* **128**, 1118–1126. https://doi.org/10.1213/ANE.0000000000003714 (2019).
38. Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R. & Singh, H. Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. *JAMA Intern. Med.* **173**, 1952–1958. https://doi.org/10.1001/jamainternmed.2013.10081 (2013).
39. Benda, N. C., Novak, L. L., Reale, C. & Ancker, J. S. Trust in AI: Why we should be designing for APPROPRIATE reliance. *J. Am. Med. Inform. Assoc.* **29**, 207–212. https://doi.org/10.1093/jamia/ocab238 (2022).
40. Bansal, G. *et al.* Updates in human–AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 2429–2437 (2019).
41. Medow, M. A., Arkes, H. R. & Shaffer, V. A. Are residents' decisions influenced more by a decision aid or a specialist's opinion? A randomized controlled trial. *J. Gen. Intern. Med.* **25**, 316–320. https://doi.org/10.1007/s11606-010-1251-y (2010).
42. Rubin, D. L. Artificial intelligence in imaging: The radiologist's role. *J. Am. Coll. Radiol.* **16**, 1309–1317. https://doi.org/10.1016/j.jacr.2019.05.036 (2019).
43. Hoffman, R. R., Mueller, S. T., Klein, G. & Litman, J. Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608 (2018).
44. Pescetelli, N., Hauperich, A.-K. & Yeung, N. Confidence, advice seeking and changes of mind in decision making. *Cognition* **215**, 104810 (2021).

45. Nguyen, A., Yosinski, J. & Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 427–436 (2015).
46. van der Waa, J., Schoonderwoerd, T., van Diggelen, J. & Neerincx, M. Interpretable confidence measures for decision support systems. *Int. J. Hum. Comput. Stud.* **144**, 102493 (2020).

Author contributions

C.R., P.C., A.C. Conceived the experiment; A.C., C.H., and the GI Genius CADx Study Group collected data; C.R., T.R., A.S., P.C., A.C. designed the experiment and methods; T.R. and A.S. performed the analysis; C.R., T.R., A.S., P.C. wrote the first draft of the manuscript; C.R., T.R., A.S., C.H., P.C., A.C. revised and edited the manuscript.

Competing interests

AC is an employee of Cosmo AI/Linkverse. CR is offering paid advice to Linkverse on a different project. The remaining authors have no conflicts of interest to disclose. We dealt with potential competing interests by pre-registering our main measures, analyses, and hypotheses before data collection (https://osf.io/y9at5). Data were analyzed by TR and AS (who are free of competing interests) and are made publicly available to the scientific community (https://osf.io/57smj). Results from planned analyses, exploratory analyses, and analyses advised by Reviewers have all been reported.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-18751-2.

Correspondence and requests for materials should be addressed to C.R. or A.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

GI Genius CADx Study Group

Giulio Antonelli[8], Halim Awadie[9], Sebastian Bernhofer[10], Sabela Carballal[11], Mário Dinis-Ribeiro[12], Agnès Fernández-Clotett[11], Glòria Fernández Esparrach[11], Ian Gralnek[9], Yuta Higasa[13], Taku Hirabayashi[14], Tatsuki Hirai[15], Mineo Iwatate[16], Miki Kawano[17], Markus Mader[10], Andreas Maieron[10], Sebastian Mattes[10], Tastuya Nakai[18], Ingrid Ordas[11], Raquel Ortigão[12], Oswaldo Ortiz Zúñiga[11], Maria Pellisé[11], Cláudia Pinto[12], Florian Riedl[10], Ariadna Sánchez[11], Emanuel Steiner[10] & Yukari Tanaka[14]

[8]Gastroenterology and Digestive Endoscopy Unit, Ospedale dei Castelli (N.O.C.), Ariccia, Italy. [9]Gastrointestinal and Liver Institute, Emek Medical Center, Afula, Israel. [10]Gastroenterology and Hepatology and Rheumatology, University Hospital of St. Pölten, St. Pölten, Austria. [11]Gastroenterology Department, Hospital Clinic of Barcelona, Barcelona, Spain. [12]Gastroenterology Department, Portuguese Oncology Institute of Porto, Porto, Portugal. [13]Department of Gastroenterology, Kita-Harima Medical Center, Ono, Japan. [14]Gastroenterology Department, Hyogo Cancer Center, Hyogo, Japan. [15]Gastroenterology Department, Sugita Genpaku Memorial Obama Municipal Hospital, Obama, Japan. [16]Gastrointestinal Center, Sano Hospital, Hyogo, Japan. [17]Kobe Red Cross Hospital, Hyogo, Japan. [18]Kobe University Hospital, Hyogo, Japan.