Gastroenterología y Hepatología 47 (2024) 481−490

Gastroenterología y Hepatología

www.elsevier.es/gastroenterologia

Gastroenterología y Hepatología

Publicación oficial de la Asociación Española para el Estudio del Hígado, Asociación Española de Gastroenterología y Grupo Español de Trabajo en Enfermedad de Crohn y Colitis Ulcerosa

Volumen 41, Número 2, Febrero 2018

Nuevas terapias en la EHNA

Manejo de aspectos psicológicos en EII: recomendación de GETECCU y ACCU

Prevalencia de insuficiencia pancreática exocrina en pancreatitis crónica

ORIGINAL ARTICLE

Design and validation of an artificial intelligence system to detect the quality of colon cleansing before colonoscopy[☆]

Antonio Z. Gimeno-García[a,*,◊], Silvia Alayón-Miranda[b,◊], Federica Benítez-Zafra[a], Domingo Hernández-Negrín[a], David Nicolás-Pérez[a], Claudia Pérez Cabañas[a], Rosa Delgado[a], Rocío del-Castillo[a], Ana Romero[a], Zaida Adrián[a], Ana Cubas[a], Yanira González-Méndez[a], Alejandro Jiménez[c], Marco A. Navarro-Dávila[d], Manuel Hernández-Guerra[a]

[a] *Gastroenterology Department, Hospital Universitario de Canarias, Instituto Universitario de Tecnologías Biomédicas (ITB) & Centro de Investigación Biomédica de Canarias (CIBICAN), Internal Medicine Department, Universidad de La Laguna, Tenerife, Spain*

[b] *Department of Computer Science and Systems Engineering, Universidad de La Laguna, Tenerife, Spain*

[c] *Research Unit, Hospital Universitario de Canarias, Tenerife, Spain*

[d] *Pharmacy Department, Hospital Universitario de Canarias, Tenerife, Spain*

Received 12 September 2023; accepted 18 December 2023
Available online 26 December 2023

KEYWORDS
Bowel cleansing;
Convolutional neural network;
Artificial intelligence;
Colonoscopy

Abstract

Background and aims: Patients' perception of their bowel cleansing quality may guide rescue cleansing strategies before colonoscopy. The main aim of this study was to train and validate a convolutional neural network (CNN) for classifying rectal effluent during bowel preparation intake as ''adequate'' or ''inadequate'' cleansing before colonoscopy.

Patients and methods: Patients referred for outpatient colonoscopy were asked to provide images of their rectal effluent during the bowel preparation process. The images were categorized as adequate or inadequate cleansing based on a predefined 4-picture quality scale. A total of 1203 images were collected from 660 patients. The initial dataset (799 images), was split into a training set (80%) and a validation set (20%). The second dataset (404 images) was used to develop a second test of the CNN accuracy. Afterward, CNN prediction was prospectively compared with the Boston Bowel Preparation Scale (BBPS) in 200 additional patients who provided a picture of their last rectal effluent.

Abbreviations: CNN, convolutional neural network; BBPS, Boston Bowel Preparation Scale; CRC, colorectal cancer; CADe, computer-assisted lesion detection; CADx, computer-assisted diagnosis; AI, artificial intelligence; DL, deep learning; LRD, low-residue diet.

[☆] *Registration number of the trial*: NCT05553977.

[*] Corresponding author.
E-mail address: agimenog@ull.edu.es (A.Z. Gimeno-García).

[◊] These authors contributed equally to this manuscript.

https://doi.org/10.1016/j.gastrohep.2023.12.009

Results: On the initial dataset, a global accuracy of 97.49%, a sensitivity of 98.17% and a specificity of 96.66% were obtained using the CNN model. On the second dataset, an accuracy of 95%, a sensitivity of 99.60% and a specificity of 87.41% were obtained. The results from the CNN model were significantly associated with those from the BBPS ($P < 0.001$), and 77.78% of the patients with poor bowel preparation were correctly classified.
Conclusion: The designed CNN is capable of classifying ''adequate cleansing'' and ''inadequate cleansing'' images with high accuracy.

PALABRAS CLAVE
limpieza de colon;
Red neuronal convolucional;
Inteligencia artificial;
Colonoscopia

Diseño y validación de un sistema de inteligencia artificial para detectar la calidad de la limpieza del colon previa a la realización de la colonoscopia

Resumen
Antecedentes y objetivos: La percepción de los pacientes sobre la calidad de su limpieza intestinal puede guiar las estrategias de limpieza de rescate antes de una colonoscopia. El objetivo principal de este estudio fue entrenar y validar una red neuronal convolucional (CNN) para clasificar el efluente rectal durante la preparación intestinal como «adecuado» o «inadecuado».
Pacientes y métodos: Pacientes no seleccionados proporcionaron imágenes del efluente rectal durante el proceso de preparación intestinal. Las imágenes fueron categorizadas como una limpieza adecuada o inadecuada según una escala de calidad de 4 imágenes predefinida. Se recopilaron un total de 1.203 imágenes de 660 pacientes. El conjunto de datos inicial (799 imágenes) se dividió en un conjunto de entrenamiento (80%) y un conjunto de validación (20%). Un segundo conjunto de datos (404 imágenes) se utilizó para evaluar la precisión de la CNN. Posteriormente, la predicción de la CNN se comparó prospectivamente con la escala de preparación colónica de Boston (BBPS) en 200 pacientes que proporcionaron una imagen de su último efluente rectal.
Resultados: En el conjunto de datos inicial, la precisión global fue del 97,49%, la sensibilidad del 98,17% y la especificidad del 96,66%. En el segundo conjunto de datos, se obtuvo una precisión del 95%, una sensibilidad del 99,60% y una especificidad del 87,41%. Los resultados del modelo de CNN se asociaron significativamente con la escala de preparación colónica de Boston ($p < 0,001$), y el 77,78% de los pacientes con una preparación intestinal deficiente fueron clasificados correctamente.
Conclusión: La CNN diseñada es capaz de clasificar imágenes de «limpieza adecuada» y «limpieza inadecuada» con alta precisión.

Introduction

A colonoscopy is the gold standard for the examination of the colon and detection of colorectal neoplastic lesions, and it has been shown to decrease the incidence and mortality rates of colorectal cancer (CRC) in a screening setting.[1,2] However, its efficiency depends on a number of factors. Quality benchmarks have been proposed by different societies to increase the efficiency of the procedure.[3,4] Cleansing quality is of paramount importance to increase the efficiency of the procedure and is closely related to the most important factors.[5,6] Despite the importance of proper bowel cleansing, the rate of colonoscopies with inadequate bowel cleansing in endoscopy units ranged from 6.8% to 33% across studies,[7,8] while a percentage between 10 and 15% is considered to be admissible by scientific societies.[3,4] Several factors have been associated with poor bowel cleansing, including those depending on the patient, staff, and those related to the institution bowel preparation protocols.[8–10] A patient's perception of his last bowel movement before the colonoscopy has been shown to be a powerful predictor of bowel cleansing ratings during colonoscopy.[11,12] A recent study carried out by our research group in a large sample of patients showed moderate agreement with the bowel cleansing rating during colonoscopy.[13] In addition, good agreement was found when the staff perception and patient perception of the last bowel movement were compared. These findings offer an excellent opportunity to test rescue cleansing interventions on the same day of the examination before the colonoscopy.

In recent years, substantial breakthroughs in several disciplines have been made using artificial intelligence (AI) applications. Machine learning, a field included in AI, refers to the development of algorithms with the ability to learn and perform certain tasks.[14] Deep learning (DL), a subset of machine learning, is based on the use of neural networks with a large number of layers and parameters.[15] These biologically inspired computational models can exceed the performance of previous forms of AI. One of the most popular forms of DL architectures is the convolutional neural network (CNN), which is specially designed to process images.[16]

In the endoscopy setting, AI computer vision applications have been stated as a research priority.[17] The two main AI system categories are computer-assisted lesion detection (CADe) and computer-assisted diagnosis (CADx).[18] However, tools for bowel preparation aid could be another field of expansion of AI in endoscopy.

The aim of this study was to design a CNN capable of automatically predicting the quality of patient cleansing at home after the intake of the bowel cleansing solution and before undergoing the colonoscopy. To achieve this, the CNN was trained on rectal effluent images acquired by the patients themselves during the bowel preparation process.

In this work, a complete description of the developed study, which involves the collection of patients' images and their subsequent labeling by medical experts, is presented. Regarding the technical aspects of this work, the chosen CNN model is presented, and its tuning and validation processes are described. Finally, the results of the CNN classification are analyzed, and the conclusions and future lines of work are described.

Material and methods

Design and setting

This study was nested in an observational prospective study conducted at the Open Access Endoscopy Unit of the Hospital Universitario de Canarias between February 2021 and May 2021 (NCT04702646).[13]

In brief, a total of 633 consecutive outpatients with a scheduled colonoscopy participated in this study. The aim of the study was to assess the agreement between the effluent characteristics of the last self-reported bowel movement by the patient and a validated bowel preparation scale (Boston Bowel Preparation Scale, BBPS).[19] Four pictures of effluents with different cleansing qualities were provided to the patient at the endoscopy unit entrance and before colonoscopy, and the patient pointed out the one that most resembled his last effluent (Supplementary figure 1). The four pictures were then categorized into two groups (adequate and inadequate cleansing) for the statistical analysis. The patients were also asked to provide pictures of their effluents. After this study, patients who requested an outpatient colonoscopy were asked to provide pictures of their effluents during bowel preparation intake. All the patients were advised to take pictures from the toilet bowl, with adequate lighting conditions, over a light background (preferably white) using their own cell phone. The images that were very dark (in which barely the bowl content could be seen) or not taken from the toilet bowl were discarded.

Overall, 660 patients (age 62.3 ± 12.9 years; 53.4% males) provided at least one image, and a total of 1342 images were collected during the bowel cleansing process (not necessarily from the last rectal effluent) at two different temporal moments (February 2021 to September 2021 and February 2022 to June 2022). A total of 139 images were discarded (96 were very dark and 43 were not taken from the toilet bowl). The first dataset was composed of 799 images, and the second dataset was composed of 404 images. All these images were used for training and validation of the CNN taking as a reference the 4-picture set. For the present study, the pictures were downloaded and labeled with consecutive numbers. Additionally, 200 patients voluntarily provided 1 image of their last rectal effluent between October 2022 and November 2022. These pictures were analyzed by the trained CNN and compared with the BBPS. The ethics committee approved the study protocol in December 2021 (Acta 20/2021). This study was registered at Clinicaltrials.gov (NCT05553977) in September 2022.

Grading

The images were independently assessed by three members of our staff (two experienced endoscopists and one experienced nurse of the endoscopy unit). The pictures were labeled as ''adequate cleansing'' (clear liquid, clear liquid with lumps) or ''inadequate cleansing'' (dark liquid, or dark liquid with solid particles) according to the aforementioned 4-picture set scale. Agreement was assessed among the three raters. In case of any discrepancy, the decision of cleansing quality was made by consensus and served as the reference standard.

According to the criteria of these experts, of the 799 images that made up the first dataset, 360 corresponded to ''adequate cleansing'', and 439 corresponded to ''inadequate cleansing''. The second dataset (404 images) was divided into 151 ''adequate'' and 253 ''inadequate'' cleansing images.

CNN algorithm description

A convolutional neural network (CNN) is a deep learning model inspired by the organization of the animal visual cortex and designed to iteratively learn spatial hierarchies of image features, from low- to high-level patterns.[20] This particular neural network consists of numerous convolution layers preceding subsampling (pooling) layers, while the ending layers are fully connected layers.[21] Each layer is composed of a variable number of neurons, and the relations between these neurons are regulated with weights. The CNN needs to be trained with labeled examples to be able to predict the correct class of a new input image. This training procedure is an iterative search of the weights' optimal values. The CNN model used in this work was VGG16[20] (Fig. 1).

The main goal of the CNN was to carry out automatic image classification into two categories: ''adequate cleansing'' and ''inadequate cleansing''. The first collected dataset, composed of 799 images, was split into a training set (80%) and a test set (20%). The CNN model was trained and initially validated on these datasets. The second collected dataset, composed of 404 images, was later used for developing a second validation.

The training strategy started with the pretrained VGG16 network using the weight values tuned for ImageNet.[22] ImageNet is an image dataset that has been used in the

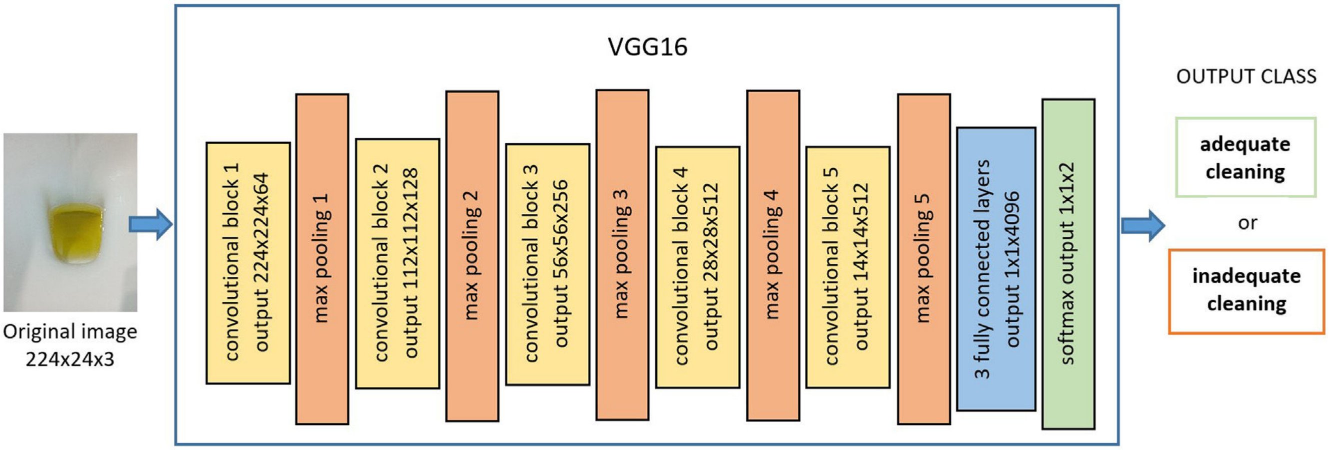

Figure 1 VGG16 architecture.

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) for image classification, a competition won using VGG16 in 2014 (Fig. 1). Starting from the pretrained network allows us to take advantage of the general learning that the network has already acquired about the most basic characteristics common to all types of images (edge detection, textures, etc.).

To transfer this learning to our classification problem, we removed the last fully connected layers (prepared for classifying 1000 classes) and attached other fully connected layers designed for classifying two classes. This methodology enables us to provide a more accurate training with the number of images collected. When the CNN-based model processes an image, it returns the probability that the image belongs to the two categories (''adequate cleansing'' and ''inadequate cleansing''). The category with the highest probability score will be considered as the CNN's predicted classification.

CNN training and evaluation

In the first training step, an iterative *fine-tuning* process was carried out, where the convolutional base network was frozen and only the newly added part was trained. In the second step, with the model obtained in the previous step, all the layers in the base network were unfrozen and iteratively tuned (*deep tuning*). Both fine tuning and deep tuning make use of the *training set* constructed from the initial dataset (80% of the whole sample). The *training set* was further divided into *training* (80%) and validation (20%) subsets. The test set (20% of the whole sample) was used for developing the initial evaluation of the final CNN model obtained during the training procedure. Since a second dataset became available later, we were able to perform a second validation study of the final CNN model. Fig. 2 shows a scheme of this procedure.

To avoid overfitting, data augmentation (an artificial procedure for increasing the dataset) consisting of random rotations, vertical and horizontal flips, and zoom was used.

The parameters related to the model's training procedure were set by trial and error. Fine and deep tuning were developed during 100 iterations (epochs) each using the RMSprop optimizer, a learning rate of 2e-5, the loss function ''categorical cross-entropy'' and a batch size of 32.

In the experiments, we used the TensorFlow (available at https://www.tensorflow.org/) and Keras (available at https://keras.io/) libraries to prepare the data and run the model. The analyses were performed with a computer equipped with a 2.9 GHz Intel Core i7-10700 processor and a Dual NVIDIA RTX3060 graphics processing unit.

After training and validation of the CNN, the CNN was used to analyze the 200 images of the last effluent provided by 200 patients. The results were then compared with the bowel preparation during the colonoscopy, rated based on the BBPS.[19] This validated scale ranges from 0 to 3 points per segment (proximal, transverse and distal colon). Bowel cleansing was adequate when each of the colon segments was scored ≥2 points. Bowel cleansing was inadequate when the score in at least one of the segments was <2 points.

Statistical analyses

The agreement among the three raters for labeling the pictures provided by the patients as ''adequate or inadequate'' cleansing was calculated by Cohen's kappa coefficient.

Statistical indicators of the trained CNN (accuracy, sensitivity, specificity, positive and negative predicted values) were also calculated.

Additionally, the prediction of the CNN (''adequate or inadequate'' cleansing) in the last 200 patients who provided images from the last rectal effluent was compared with the BBPS scores categorized as adequate preparation (BBPS score ≥2 per segment) or inadequate preparation (BBPS score <2 per segment). The agreement between the CNN prediction of bowel cleansing and the result of the BBPS score was calculated by Cohen's kappa coefficient. Statistical indicators (sensitivity, specificity, positive, negative predicted values, positive and negative likelihood ratios) were calculated. In addition, the agreement calculated by the Cohen's kappa coefficient plus 95% confidence intervals between the observers' assessment and the BBPS was calculated. Data were analyzed with the Statistical Package for Social Sciences v. 25.0 (Armonk, NY: IBM Corp).

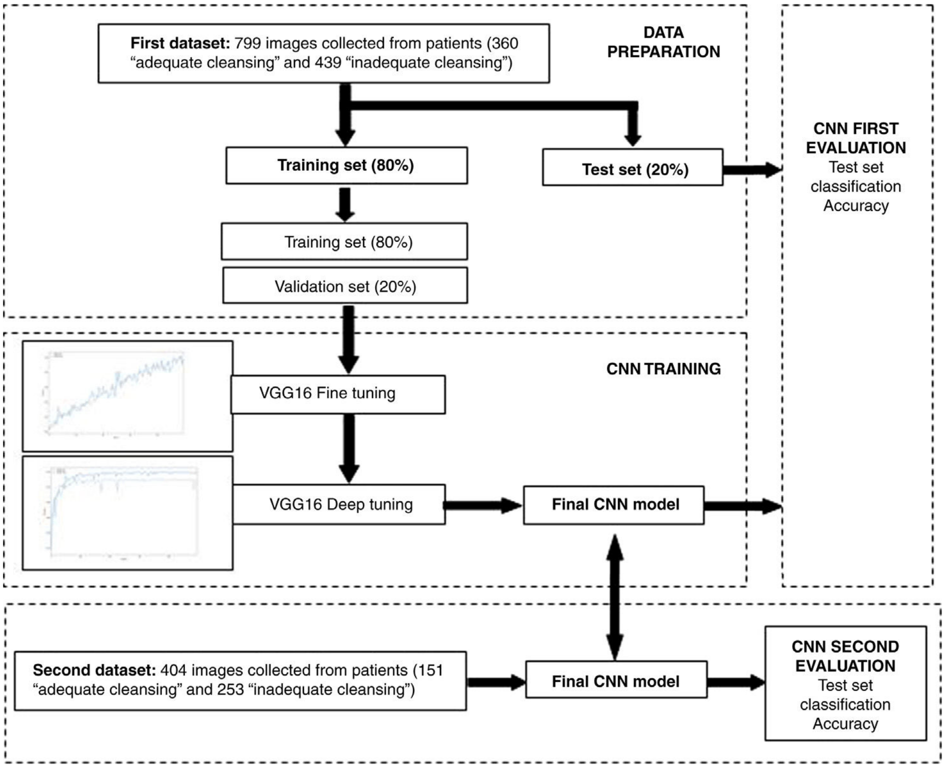

Figure 2 Flow chart of the CNN training and test procedures.

Results

For the training and validation of the CNN taking as a reference the 4-picture set, 660 patients were included (males 53.2%, age ± SD 62.7 ± 13.7 years). A total of 161 had significant comorbidities (24.4%). Overall, 50.2% were prepared with 2 l of polyethylene glycol plus ascorbic acid, 29.5% with sodium picosulphate and magnesium citrate, 10.6% with 1 l of polyethylene glycol and ascorbic acid and 9.7% with 4 l of polyethylene glycol.

VGG16-based model development and performance evaluation

Agreement was calculated among the three observers according to the 4-picture set scale. There was good agreement in cleansing quality between observers 1 and 2 ($k = 0.755$), observers 1 and 3 ($k = 0.851$) and observers 2 and 3 ($k = 0.788$).

During the fine and deep tuning procedure, it was observed that the accuracy of the model increased as the number of iterations increased and the weights were adjusted to better values. This behavior is reflected in the learning curves shown in Fig. 3.

An accuracy of 99.84%, a sensitivity of 100% and a specificity of 99.65% were obtained on the training set using the final CNN model obtained after training. On the test set, an accuracy of 88.34%, a sensitivity of 91.01% and a specificity of 85.13% were obtained.

In summary, on the initial dataset, a global accuracy of 97.49%, a sensitivity of 98.17% and a specificity of 96.66% were obtained using the CNN model. The positive and negative predictive values were 97.29% and 97.75%, respectively. The confusion matrix is shown in Table 1, and the statistical performances are shown in Table 2.

Table 1 Confusion matrix of VGG16 detection versus expert classification when classifying the initial image dataset and second image dataset.

	Expert cleansing classification	
	Adequate	Inadequate
Cleansing in the initial image dataset		
VGG16 classification		
Adequate	348	8
Inadequate	12	431
Cleansing in the second image dataset		
VGG16 classification		
Adequate	132	1
Inadequate	19	252

VGG16-based model second evaluation

To more deeply study the generalization capacity of the designed classifier, a second test of the model was carried out with new images not considered in the design process, as described in the previous section. Specifically, the 404 images contained in the second dataset were used. From this collected pool of images ($n = 404$), 151 images were labeled

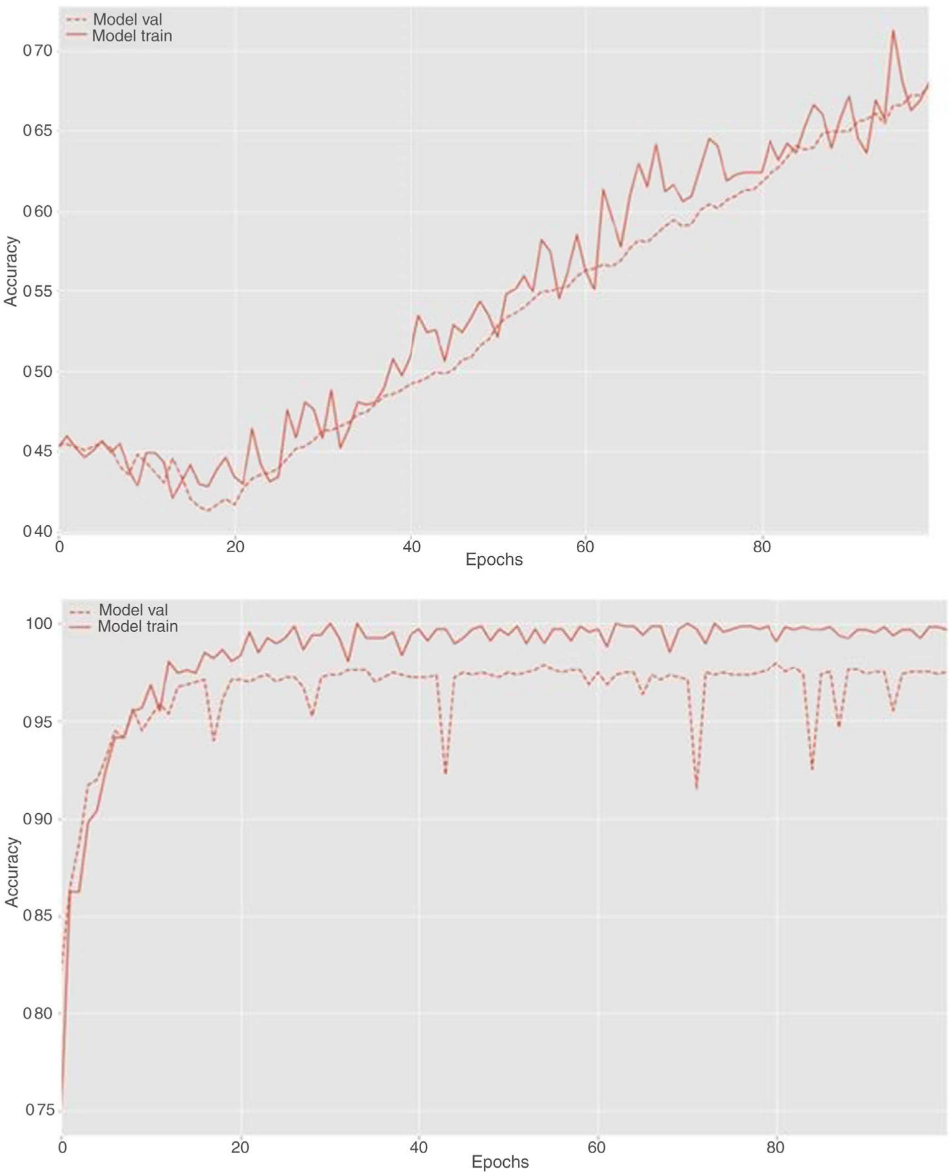

Figure 3 Learning curve for fine (up) and deep (down) tuning of the CNN-based model trained. For each iteration (epoch), the CNN classification accuracies on the training and validation sets are displayed.

Table 2 Sensitivity, specificity, and positive and negative predictive values of VGG16 detection versus expert classification when classifying the first and second image datasets.

	Image dataset	
	First	Second
Sensitivity – % (95% CI[a])	98.18 (96.45–99.07)	99.6 (97.8–99.93)
Specificity – % (95% CI)	96.67 (94.26–98.08)	87.42 (81.18–91.79)
Positive predictive value – % (95% CI)	97.29 (95.33–98.44)	92.99 (89.31–95.47)
Negative predictive value – % (95% CI)	97.75 (95.63–98.86)	99.25 (95.86–99.87)
Positive likelihood ratio (95% CI)	53 (27–105)	221 (31–1565)
Negative likelihood ratio (95% CI)	0.03 (0.02–0.06)	0.1 (0.08–0.19)

[a] 95% confidence interval.

Table 3 Sensitivity, specificity, and positive and negative predictive values of VGG16 detection versus the Boston Bowel Preparation Scale (BBPS).

Sensitivity –*n*, % (95% CI[a])	21/27, 77.78 (59.24–89.39)
Specificity – *n*, % (95% CI)	138/173, 79.77 (73.17–85.08)
Positive predictive value – *n*, % (95% CI)	21/56, 37.50 (26.01–50.59)
Negative predictive value – *n*, % (95% CI)	138/144, 95.83 (91.21–98.08)
Positive likelihood ratio (95% CI)	3.84 (2.69–5.5)
Negative likelihood ratio (95% CI)	0.28 (0.14–0.57)

Cohen's kappa coefficient (*k*) between the CNN prediction and the BBPS: $k = 0.396$, 95% CI [0.254–0.538], $P < 0.001$.

[a] 95% confidence interval.

by the experts as representing ''adequate cleansing'' and 253 images reflected ''inadequate cleansing'' preparations.

When using the obtained CNN model on this dataset, the global accuracy was 95%, with a sensitivity of 99.60% and a specificity of 87.41%. The positive and negative predictive values were 92.99% and 99.25%, respectively. The confusion matrix is shown in Table 1, and the statistical performances are shown in Table 2.

CNN prediction of cleansing quality during the colonoscopy

A total of 200 patients provided a picture of the last rectal effluent (age 60.94 ± 11.66 years; 51.5% males). The main indications were CRC screening (30.5%) and post-polypectomy surveillance (23.0%) (Supplementary Table 1). Baseline characteristics of these patients are shown in Supplementary Table 2. Overall, 58 patients (29%) had comorbidities (diabetes mellitus, cirrhosis, stroke, and significant chronic kidney disease defined as renal glomerular filtration <60 mL/min). Agreement was calculated among the three observers according to the 4-picture set scale. There was good agreement in cleansing quality between observers 1 and 2 ($k = 0.775$), observers 1 and 3 ($k = 0.748$), and observers 2 and 3 ($k = 0.851$).

Overall, 173 patients (86.5%) were rated as having adequate cleansing following the BBPS, and 27 (13.5%) as having inadequate cleansing. There was fair agreement between the global observer decision on cleansing quality and BBPS ($k = 0.377$, 95% CI [0.212–0.542], $P < 0.001$). The CNN predicted adequate bowel cleansing in 144 patients (72.0%) and inadequate cleansing in 56 patients (28.0%). The agreement between the CNN prediction and the BBPS was also fair ($k = 0.396$, 95% CI [0.253–0.539], $P < 0.001$). The sensitivity, specificity, positive predictive value, and negative predictive value of the CNN prediction compared to those of the BBPS are shown in Table 3. The CNN was capable of correctly detecting 21 out of 27 patients (77.78%) with poor bowel preparation during the colonoscopy. A total of 65 patients had colorectal adenomas (32.5%). There were no statistically significant differences between patients with poor bowel preparation based on the CNN prediction ($n = 15$, 26.8%) and patients with a predicted adequate bowel cleansing ($n = 50$, 34.7%, $P = 0.282$). Another analysis was carried out depending on the bowel solution ingested. The statistical performances are shown in Supplementary Table 3.

Discussion

In this study, we designed and validated a CNN capable of precisely discriminating different qualities of bowel cleansing according to a predetermined set of images. Most work presented in the literature are focused on the application of AI techniques to improve detection and diagnosis by colonoscopy but not to improve the cleansing quality.

Research on the improvement of colonoscopy quality factors is a timely subject. Colon cleansing is closely linked to the main outcomes, such as adenoma detection rate or interval CRC, and several studies attempting to improve colon cleansing have been reported in recent years.[5,23–26] However, little progress has been made on this issue, and in clinical practice, the rate of patients with inadequate bowel cleansing frequently exceeds the benchmark of 10%–15%.[3,26,27]

Great efforts have been made to test the impact of quantitative and qualitative changes in the type of diet (regular diet, liquid diet or low-residue diet), the number of low-residue diet (LRD) days, the amount of bowel solution ingested (low-volume or high-volume preparations), the specific type of bowel solution (isosmotic or hyperosmotic solutions with additional adjuvants, etc.), or the use of intensive preparation protocols (increasing the amount of bowel solution intake plus the number of LRD days plus adjuvants). However, in general, little benefit has been found when different preparation protocols were compared.[5,24,28–30]

Recently, in a study carried out by our group that included 633 patients in a derivation cohort and 378 patients in a validation cohort, we found that the patients' perception of their cleansing quality while ingesting the cleansing solution was a powerful predictor of colon cleansing quality assessed during the colonoscopy using a validated bowel preparation scale (Boston Bowel Preparation Scale, BBPS).[13] Although, in this study, agreement between patients' perception and BBPS was fair, in the multiple logistic regression analysis, it was the most powerful predictor of bowel cleansing during colonoscopy ahead of other well-recognized factors such as aging, the lack of adherence to bowel preparation intake or suffering from comorbidities such as diabetes mellitus. In that study, the patients pointed out the image that was most similar to their last rectal effluent during bowel preparation from a 4-cleansing quality picture set of different cleansing qualities (Supplementary figure 1). This first study encouraged us to design a CNN capable of predicting the cleansing quality of the patient after ingesting the cleansing solution based on the same 4-cleansing quality picture set.

High accuracy values were achieved using this model. However, as expected, the classification results were slightly worsened when images not considered in the training process were analyzed. Nevertheless, the accuracy was still good, which shows that the CNN model has generalization capacity.

Interestingly, slightly more errors were noted on both datasets when classifying images as ''adequate cleansing'' using the CNN model. This may be due to the presence of images that are at the limit of what is acceptable in terms of preparation quality.

To our knowledge, only two studies, both conducted in Chinese populations, have approached this problem in a similar way.[31,32] In both studies a CNN was trained with labeled pictures of rectal effluents. In the first study,[31] when the CNN prediction was compared with the BBPS, it was noted that the cleansing quality using the CNN model was not well discriminated, since only 6 out of 71 patients with a BBPS <6 points (8.45%) were correctly classified; the CNN incorrectly classified the remaining images as adequate preparation. Conversely, 26 patients with a BBPS $\geq$6 points were incorrectly classified as having inadequate preparation. In light of these findings, this CNN would probably have a residual clinical impact for guiding rescue strategies in patients with a CNN prediction of poor cleansing quality the same day of the colonoscopy appointment to avoid repeating the procedure.

Regarding the training process of the CNN in the abovementioned study, although the authors reported an accuracy of 97%, detailed information about the training process is missing in their study. In the second study,[32] carried out in 524 patients, a smartphone application driven by a CNN showed an improvement in the quality of colonic cleansing compared to a control group. However, the rate of adequate colon cleansing in the control group was much lower than expected (88.54% in the AI group vs 65.59% in the control group). Regarding the CNN, accuracy was 95.15% in the test dataset, similar to our study. However, the CNN and the training used in this study were different from the ones used in our study.

Our CNN was trained on pictures of rectal effluents labeled by three raters following a 4-cleansing quality picture set, which has been proven to be in acceptable agreement with the BBPS, as described in a previous study.[13] Although, it was not the main aim of the present study, the overall performance for predicting adequate or inadequate cleansing following the BBPS seems to be acceptable. Our CNN was able to detect 21 out of the 27 patients with inadequate preparation. Since the BBPS score is determined after colon washing and suctioning, we hypothesized that this maneuver could have rescued some patients who otherwise would have inadequate preparation.

We believe that the promising results of our study could be the basis for designing mobile applications that could prove the preparation quality of the colon before coming to an appointment for colonoscopy and launch rescue cleansing strategies in a timely manner, decreasing repeated procedures, inconvenience for patients and costs.

Our study has some limitations. First, although our results are promising, the datasets of images used for training and validation of this study were not necessarily taken from the last bowel movement but from any rectal effluent during the bowel preparation process. In addition, the proportion of images labeled as adequate and inadequate cleansing during the training of the CNN do not represent that in clinical practice because the images were not limited to the last rectal effluent. However, in this study, it was necessary to obtain the largest possible number and variety of pictures. Although, the number of images for the CNN training could seem small, the use of techniques such as transfer learning greatly reduce the number of images needed for this task. This is proven by the fine and deep tuning learning curves, both showing that the CNN learns reaching a great accuracy without overfitting. Second, although we compared the prediction of the CNN with a validated colon cleansing scale during colonoscopy, patients voluntarily provided the pictures and therefore are not representative of our population. In addition, we are aware that the number of patients who provided a picture of only their last rectal effluent was limited. However, our agreement results between the CNN prediction and the BBPS ($k = 0.396$) are comparable to those obtained in our previous study ($k = 0.374$),[13] which makes them reliable. Third, we did not identify significant differences between patients with adequate and poor bowel cleansing predictions in terms of adenoma detection rates. However, it is essential to acknowledge that the primary objective of the study was not focused on this aspect. Adequately designed studies to address this objective would be of interest in order to generalize the use of this application.

Finally, although the sensitivity and specificity of the CNN can be improved, it does present a high NPV (>95%), allowing for optimal exclusion of patients with inadequate cleansing. However, we acknowledge that the PPV is improvable, and patients with inadequate cleansing are not detected with high precision. Therefore, if additional colonic preparation were administered to patients with an inadequate prediction, a non-negligible proportion of patients would be over-treated. Nevertheless, we believe it would be worthwhile in order to reduce the percentage of inadequate colonoscopies that would need to be repeated.

In conclusion, the developed AI tool, as described in this work, is capable of classifying ''adequate cleansing'' and ''inadequate cleansing'' images with high accuracy. Studies in clinical practice are warranted to assess the guidance of proper bowel preparation as a rescue strategy before the colonoscopy appointment.

Authors' contributions

Conception and design: Antonio Zebenzuy Gimeno-García, Federica Benítez-Zafra.

Performed procedures: Antonio Zebenzuy Gimeno-García, David Nicolás Pérez, Manuel Hernández-Guerra, Zaida Adrián, Yanira González-Méndez, Silvia Alayón Miranda.

Data collection: Domingo Hernández-Negrín, Rocío del Castillo, Alvaro Peralta, Ana Romero, Ana Cubas, Claudia Pérez.

Data analysis and interpretation: Alejandro Jiménez, Antonio Zebenzuy Gimeno-García.

Drafting of the manuscript: Antonio Zebenzuy Gimeno-García, Marco Antonio Navarro-Dávila.

Critical revision: Manuel Hernández-Guerra, Silvia Alayón Miranda.

Final approval of the article: Antonio Zebenzuy Gimeno-García.

Ethical approval

The authors declare that all the participants signed an informed consent. The protocol was approved by the local ethics committee of Hospital Universitario de Canarias 21/12/2021 (Acta 20/2021).

Funding

This study was supported by a grant from Fundación Canaria Instituto de Investigación Sanitaria de Canarias (FIISC) (PIFI-ISC20/04). This foundation was not involved in any phase of this research.

Conflicts of interest

Dr. AZ Gimeno-García has received a research grant from Fundación Canaria Instituto de Investigación Sanitaria de Canarias (FIISC). The other authors do not have conflicts of interest.

Acknowledgments

The authors want to thank Jorge Mon, Yaiza Cedrés, Eduardo Quintana, Silvia Blanco, Maribel Jorge, Rosa Delgado, Gloria Lorenzo, Ana Cubas for their support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gastrohep.2023.12.009.

References

1. Atkin W, Wooldrage K, Parkin DM, Kralj-Hans I, MacRae E, Shah U, et al. Long term effects of once-only flexible sigmoidoscopy screening after 17 years of follow-up: the UK Flexible Sigmoidoscopy Screening randomised controlled trial. Lancet. 2017;389:1299–311, http://dx.doi.org/10.1016/S0140-6736(17)30396-3.
2. Pilonis ND, Bugajski M, Wieszczy P, Rupinski M, Pisera M, Pawlak E, et al. Participation in competing strategies for colorectal cancer screening: a randomized health services study (PICCOLINO Study). Gastroenterology. 2021;160:1097–105, http://dx.doi.org/10.1053/j.gastro.2020.11.049.
3. Kaminski MF, Thomas-Gibson S, Bugajski M, Bretthauer M, Rees CJ, Dekker E, et al. Performance measures for lower gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. Endoscopy. 2017;49:378–97, http://dx.doi.org/10.1055/s-0043-103411.
4. Rex DK, Schoenfeld PS, Cohen J, Pike IM, Adler DG, Fennerty MB, et al. Quality indicators for colonoscopy. Gastrointest Endosc. 2015;81:31–53, http://dx.doi.org/10.1016/j.gie.2014.07.058.
5. Hassan C, East J, Radaelli F, Spada C, Benamouzig R, Bisschops R, et al. Bowel preparation for colonoscopy: European Society of Gastrointestinal Endoscopy (ESGE) Guideline – update 2019. Endoscopy. 2019;51:775–94, http://dx.doi.org/10.1055/a-0959-0505.
6. Robertson DJ, Lee JK, Boland CR, Dominitz JA, Giardiello FM, Johnson DA, et al. Recommendations on fecal immunochemical testing to screen for colorectal neoplasia: a consensus statement by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology. 2017;152, http://dx.doi.org/10.1053/j.gastro.2016.08.053, 1217–1237.e3.
7. Adams WJ, Meagher AP, Lubowski DZ, King DW. Bisacodyl reduces the volume of polyethylene glycol solution required for bowel preparation. Dis Colon Rectum. 1994;37:229–33, http://dx.doi.org/10.1007/BF02048160 [discussion 233–224].
8. Hassan C, Fuccio L, Bruno M, Pagano N, Spada C, Carrara S, et al. A predictive model identifies patients most likely to have inadequate bowel preparation for colonoscopy. Clin Gastroenterol Hepatol. 2012;10:501–6, http://dx.doi.org/10.1016/j.cgh.2011.12.037.
9. Dik VK, Moons LM, Huyuk M, van der Schaar P, de Vos Tot Nederveen Cappel WH, Ter Borg PC, et al. Predicting inadequate bowel preparation for colonoscopy in participants receiving split-dose bowel preparation: development and validation of a prediction score. Gastrointest Endosc. 2015;81:665–72, http://dx.doi.org/10.1016/j.gie.2014.09.066.
10. Gimeno-Garcia AZ, Baute JL, Hernandez G, Morales D, Gonzalez-Perez CD, Nicolas-Perez D, et al. Risk factors for inadequate bowel preparation: a validated predictive score. Endoscopy. 2017;49:536–43, http://dx.doi.org/10.1055/s-0043-101683.
11. Fatima H, Johnson CS, Rex DK. Patients' description of rectal effluent and quality of bowel preparation at colonoscopy. Gastrointest Endosc. 2010;71, http://dx.doi.org/10.1016/j.gie.2009.11.053, 1244–1252.e2.
12. Harewood GC, Wright CA, Baron TH. Assessment of patients' perceptions of bowel preparation quality at colonoscopy. Am J Gastroenterol. 2004;99:839–43, http://dx.doi.org/10.1111/j.1572-0241.2004.04176.x.
13. Gimeno-Garcia AZ, Benitez-Zafra F, Hernandez A, Hernandez-Negrin D, Nicolas-Perez D, Hernandez G, et al. Agreement between the perception of colon cleansing reported by patients and colon cleansing assessed by a validated colon cleansing scale. Gastroenterol Hepatol. 2023, http://dx.doi.org/10.1016/j.gastrohep.2023.02.009.
14. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380:1347–58, http://dx.doi.org/10.1056/NEJMra1814259.
15. Shinde PPSS. A review of machine learning and deep learning applications. In: Fourth international conference on computing communication control and automation (ICCUBEA). 2018. p. 1–6.
16. O'Shea K, Nash R. An introduction to convolutional neural networks; 2015 arXiv:1511.08458.
17. Berzin TM, Parasa S, Wallace MB, Gross SA, Repici A, Sharma P. Position statement on priorities for artificial intelligence in GI endoscopy: a report by the ASGE Task Force. Gastrointest Endosc. 2020;92:951–9, http://dx.doi.org/10.1016/j.gie.2020.06.035.
18. Mori Y, Misawa M, Kudo SE. Challenges in artificial intelligence for polyp detection. Dig Endosc. 2022;34:870–1, http://dx.doi.org/10.1111/den.14279.
19. Lai EJ, Calderwood AH, Doros G, Fix OK, Jacobson BC. The Boston bowel preparation scale: a valid and reliable instrument for colonoscopy-oriented research. Gastrointest Endosc. 2009;69 Pt 2:620–5, http://dx.doi.org/10.1016/j.gie.2008.05.057.
20. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts CNN architectures, challenges, applications, future directions. J Big Data. 2021;8:53, http://dx.doi.org/10.1186/s40537-021-00444-8.
21. Takhur R. Step by step VGG16 implementation in Keras for beginners; 2019 https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-for-beginners-a833c686ae6c
22. https://www.image-net.org/download.php.IdAa.
23. Cho JH, Goo EJ, Kim KO, Lee SH, Jang BI, Kim TN. Efficacy of 0.5-L vs 1-L polyethylene glycol containing ascorbic acid as additional colon cleansing methods for inadequate bowel preparation as expected by last stool examina-

tion before colonoscopy. World J Clin Cases. 2019;7:39–48, http://dx.doi.org/10.12998/wjcc.v7.i1.39.

24. Gimeno-Garcia AZ, de la Barreda Heuser R, Reygosa C, Hernandez A, Mascareno I, Nicolas-Perez D, et al. Impact of a 1-day versus 3-day low-residue diet on bowel cleansing quality before colonoscopy: a randomized controlled trial. Endoscopy. 2019;51:628–36, http://dx.doi.org/10.1055/a-0864-1942.
25. Gimeno-Garcia AZ, Hernandez G, Aldea A, Nicolas-Perez D, Jimenez A, Carrillo M, et al. Comparison of two intensive bowel cleansing regimens in patients with previous poor bowel preparation: a randomized controlled study. Am J Gastroenterol. 2017;112:951–8, http://dx.doi.org/10.1038/ajg.201753.
26. Johnson DA, Barkun AN, Cohen LB, Dominitz JA, Kaltenbach T, Martel M, et al. Optimizing adequacy of bowel cleansing for colonoscopy: recommendations from the US multi-society task force on colorectal cancer. Gastroenterology. 2014;147:903–24, http://dx.doi.org/10.1053/j.gastro.2014.07.002.
27. Alvarez-Gonzalez MA, Pantaleon Sanchez MA, Bernad Cabredo B, Garcia-Rodriguez A, Frago Larramona S, Nogales O, et al. Educational nurse-led telephone intervention shortly before colonoscopy as a salvage strategy after previous bowel preparation failure: a multicenter randomized trial. Endoscopy. 2020;52:1026–35, http://dx.doi.org/10.1055/a-1178-9844.
28. Ahumada C, Pereyra L, Galvarini M, Mella J, Gomez E, Pedreira SC, et al. Efficacy and tolerability of a low-residue diet for bowel preparation: systematic review and meta-analysis. Surg Endosc. 2022;36:3858–75, http://dx.doi.org/10.1007/s00464-021-08703-8.
29. Machlab S, Martinez-Bauer E, Lopez P, Pique N, Puig-Divi V, Junquera F, et al. Comparable quality of bowel preparation with single-day versus three-day low-residue diet: randomized controlled trial. Dig Endosc. 2021;33:797–806, http://dx.doi.org/10.1111/den.13860.
30. Xie Q, Chen L, Zhao F, Zhou X, Huang P, Zhang L, et al. A meta-analysis of randomized controlled trials of low-volume polyethylene glycol plus ascorbic acid versus standard-volume polyethylene glycol solution as bowel preparations for colonoscopy. PLoS One. 2014;9:e99092, http://dx.doi.org/10.1371/journal.pone.0099092.
31. Lu YB, Lu SC, Huang YN, Cai ST, Le PH, Hsu FY, et al. A novel convolutional neural network model as an alternative approach to bowel preparation evaluation before colonoscopy in the COVID-19 era: a multicenter, single-blinded, randomized study. Am J Gastroenterol. 2022;117:1437–43, http://dx.doi.org/10.14309/ajg.0000000000001900.
32. Zhu Y, Zhang DF, Wu HL, Fu PY, Feng L, Zhuang K, et al. Improving bowel preparation for colonoscopy with a smartphone application driven by artificial intelligence. NPJ Digit Med. 2023;6:41, http://dx.doi.org/10.1038/s41746-023-00786-y.