

Diagnostic accuracy of ChatGPT-4 in light of data derived from patient's clinical and imaging findings in gastrointestinal surgery

✉ Suleyman ORMAN ¹

¹ Department of Gastroenterology Surgeon, Bursa Yuksek Ihtisas Research and Training Hospital, Bursa, Turkey

Abstract

Aim: This study aims to evaluate the diagnostic accuracy (DA) of ChatGPT-4 in GIS applications.

Methods: 231 GIS cases from between October 1 2021 and January 10 2024 were sourced from the International Journal of Surgery Case Reports. ChatGPT-4 was used to generate diagnoses based on the patient's medical histories and imaging findings. Each case was categorized by anatomical location (esophagus, gastric, small intestine, and colon), and divided into emergency and elective groups.

Results: ChatGPT-4 generated 2-10 differential diagnoses per case (median: 5, mean: 4.8 ± 1.2). No significant association was found between age, gender, or the number of differential diagnoses ($p=0.687$ and $p=0.862$). A significant difference was found between anatomical localizations according to the urgency of the cases ($p<0.001$). Overall, ChatGPT-4 achieved 83.1% accuracy for differential diagnoses and 57.1% for final diagnoses ($p<0.001$ and $p=0.035$). During the trained period, accuracy was 82.9% (differential) and 59.1% (final), while in the non-trained period it was 83.6% (differential) and 52.2% (final).

Conclusion: In this large-scale, case-based evaluation, ChatGPT-4 demonstrated a substantially higher accuracy in generating differential diagnoses (DA 83.1%) compared to establishing final diagnoses (DA 57.1%). Diagnostic accuracy was slightly improved during the trained period, particularly for final diagnoses (59.1% vs. 52.2%). These findings indicate that ChatGPT-4 may serve as a valuable clinical decision support tool in the early diagnostic stage by broadening the differential diagnosis list, while its limited accuracy in final diagnosis highlights the need for cautious integration into practice.

Keywords: ChatGPT-4; Gastrointestinal Surgery; Diagnostic Accuracy; Artificial Intelligence; Clinical Decision Support

Correspondence: Assoc. Prof. MD, Department of Gastroenterology Surgeon, Bursa Yüksek İhtisas Research and Training Hospital, Bursa, Turkey

E-mail: suleymanorm@hotmail.com ORCID-ID: orcid.org/0000-0003-4840-1279

Received: 08.07.2025 Accepted: 29.09.2025

Cite this article as: Orman S. Diagnostic accuracy of ChatGPT-4 in light of data derived from patient's clinical and imaging findings in gastrointestinal surgery EurJHum Health.2025;3(2):28-42.

©Copyright 2025 by the European Journal of Human Health.

Licensed by Creative Commons Attribution - Non Commercial - No Derivatives (CCBY-NC-ND) 4.0 International License.



Introduction

The rapid development of artificial intelligence (AI) and large language models (LLMs) has led to the emergence of advanced chatbots such as ChatGPT-4, which can generate human-like responses and provide support across various domains, including healthcare [1-4]. Increasingly, medical disciplines such as radiology, oncology, and ophthalmology are exploring the clinical applications of ChatGPT-4, particularly in decision-making and diagnostic support [5-10].

In gastrointestinal surgery (GIS), AI systems have shown promise in enhancing the detection and diagnosis of gastrointestinal diseases. Previous studies demonstrated improved sensitivity and specificity of AI-assisted approaches in identifying conditions such as esophageal squamous cell carcinoma, Barrett's esophagus, gastric cancer, inflammatory bowel disease, and colorectal cancer [11-16]. Integrating AI with endoscopist assessment—through radiomics, imaging modalities, and histological evaluation—has been associated with greater diagnostic accuracy [17-19].

However, prior research exploring the role of ChatGPT in gastroenterology has remained limited in scope. For example, Horiuchi et al. [20] and Ueda et al. [21] primarily assessed ChatGPT's responses in hypothetical or simulated clinical scenarios rather than in real patient datasets, while Wei et al. focused on systematic reviews and meta-analyses of ChatGPT's general medical performance [2]. These studies provided valuable preliminary insights but did not evaluate the tool in the context of authentic, case-based gastrointestinal surgery (GIS) decision-making.

Therefore, the novelty of the present study lies in its use of real-world patient cases incorporating both clinical and radiological data to examine the diagnostic accuracy of ChatGPT-4 across diverse GIS conditions. By directly comparing model outputs with established clinical diagnoses, this study aims to provide a more comprehensive and practice-oriented evaluation of ChatGPT-4's potential role in gastrointestinal surgery.

Methods

Study Design

We assessed the diagnostic capability of ChatGPT-4, assuming that the presented clinical and imaging findings were valid and objectively produced via previously published real-life cases. All data that was derived and utilized from the cases of International Journal of Surgery Case Reports' open access sources that were published between October 1, 2021, and January 10, 2024. The findings obtained from case images and information was input following the precedent set by previous studies [20, 22, 23]. Based on this data, ChatGPT-4 generated differential and final diagnoses. We then interpreted the rate of correct diagnoses as an estimate of DA. This study was designed following the Standards for Reporting Diagnostic Accuracy Studies (STARD) statement. Ethics committee approval was not required as this study only utilized previously published real-life cases. All data were fully anonymized and publicly available, and no direct patient involvement occurred. Therefore, the study did not raise additional ethical concerns or require further institutional approval.

Setting

The International Journal of Surgery Case Reports is a peer-reviewed monthly journal that presents surgical cases taken from all fields of surgery. ChatGPT-4 was initially trained on data available up to September 2021 and lately has been trained on data up to April 2023. It was launched on March 14, 2023, for public use [1, 10, 24]. Therefore, diagnostic cases presented before September 2021 were not utilized in the study because of potential bias [24]. For this reason, 231 GIS-related consecutive cases between October 1, 2021 to January 10, 2024 were included. Given that ChatGPT-4 was last trained in April 2023, the data collection from the non-training period was included to evaluate the enhanced diagnostic accuracy of ChatGPT-4. We extracted the medical history and imaging findings of selected patients from related sections of each published study (Figure 1).

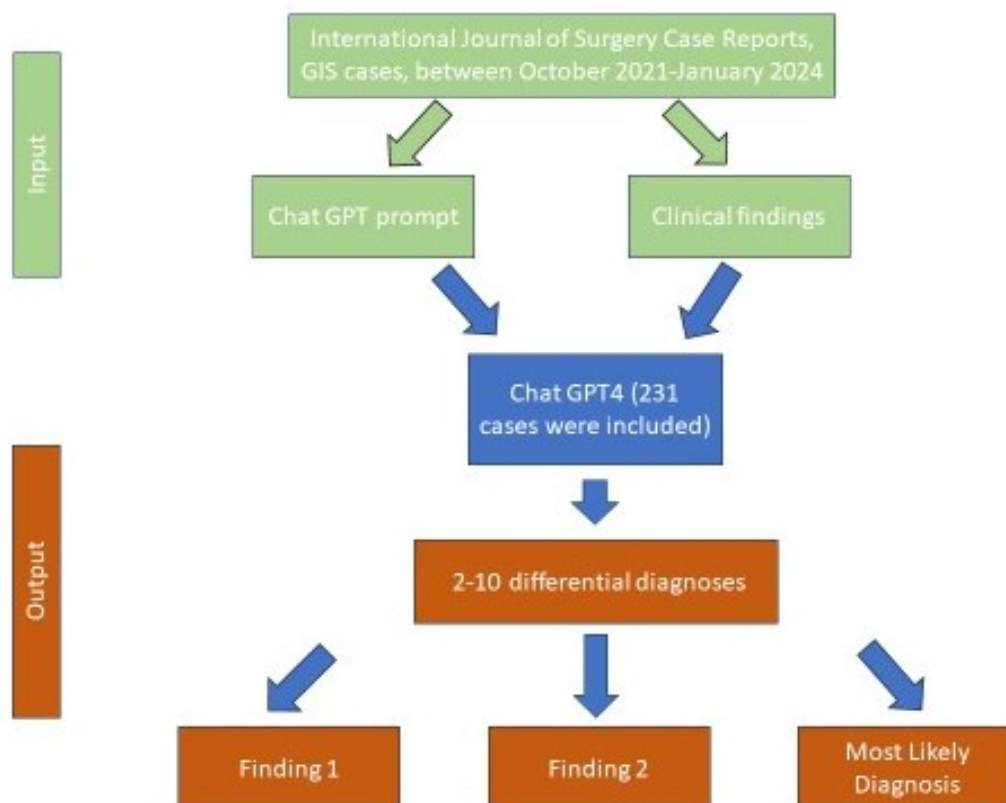


Figure 1. Extraction of patients' medical histories and imaging findings from published studies.

Eligibility criteria

Each case was categorized based on anatomical location, such as esophagus, gastric, small intestine or colon [25]. A distinction was made between emergency and elective cases. During case identification, the type of hospital applied to, major complaints, duration of complaints, and relevant medical history for the previous months were included. Vital signs at the time of application, physical examination findings, and laboratory findings, if any, were entered into ChatGPT. In addition, a basic summary of findings from the radiological images was uploaded to the program. These categorizations and case selections were made and reviewed by both a gastrointestinal surgeon and a radiologist. The gastrointestinal surgeon checked whether the anamnesis and symptoms were appropriately mentioned in each case. The radiologist verified the appropriateness of the presented imaging data and elucidated any radiologically significant

findings from the images. At this point, all findings were input to ChatGPT-4. Exclusions from the study comprised those cases that were not solely related to diagnosis issues or lacked detailed radiological evaluation, any diseases of the hepato-pancreato-biliary system that were associated with entirely different pathophysiological backgrounds; any genetic studies, and cases only related to issues of surgical technique or concerning inguinal hernia-related conditions with obvious findings. To preserve homogeneity, it was considered more appropriate to focus on diseases within the luminal regions of the gastrointestinal tract, which share similar pathophysiological disorders and approximate histological substructures. Due to the similar infrastructure of organs with lumen in GIS, these organs and related disorders were included in the scope of this study. Whole cases were mentioned in Tables 1 and Table 2 as 'Cases According to Diagnosis'. Therefore, due to this comparable infrastructure, the scope of this study encompassed GIS organs with lumen and their related disorders [11, 24].

Table 1: Cases According to Esophageal and Stomach Diagnoses

ESOPHAGEAL CASES (n= 29)			STOMACH CASES (n=23)		
Type of Presentation	Cases with Main Cause	n	Type of Presentation	Cases with Main Cause	n
Obstruction- Dysphagia (n=17)	Achalasia	4	Obstruction (n=11)	Gastric Adenocancer	2
	Esophageal Adenocarcinoma	2		Bezoars	2
	Leiomyoma	2		Liposarcoma	1
	Esophageal Diverticula	2		Volvulus	1
	Foreign Body Retantion	2		Heterotopic Pancreas	1
	GIST	1		Aortofemoral Bypass Graft	1
	Mucocele	1		Sleeve Gastrectomy	1
	Duplication Cyst	1		Gastric Pouch Enlargement	1
	Retention Cyst	1		Bouveret’s Syndrome	1
	Congenital Stenosis	1			
Perforation (n=11)	Boerhaave’s Syndrome	3	Bleeding (n=4)	GIST	2
	Hiatal Hernia	2		Dieulafoy’s Lesion	1
	Foreign Body	2	Perforation (n=3)	Schwannoma	1
	Electrocautery Injury	1		Peptic Ulcer	2
	Esophageal Adenocarcinoma	1	Lymphoma	1	
	Penetrating Injury	1	Dizziness (n=1)	GIST	1
	Achalasia	1	Bone Metastasis (n=1)	Gastric Adenocarcinoma	1
			Hepatic Metastasis (n=1)	Gastric Neuroendocrine Tumor	1
Incidental (n=1)	Granular Cell Tumor	1	Amnesia (n=1)	Gastric Adenocarcinoma	1
			Weight Loss (n=1)	Breast Cancer Metastasis	1
Total number of cases n=52					

Table 2: Cases According to Small Intestine and Colorectal Diagnoses

SMALL INTESTINE CASES (n=90)			COLORECTAL CASES (n=89)		
Type of	Cases with Main Cause	n	Type of	Cases with Main Cause	n
Obstruction (n=56)	Intussusception	10	Obstruction (n=43)	Intussusception	15
	Meckel’s Diverticulum	9		Volvulus	9
	Gallstone Ileus	6		Diverticulitis	2
	Benign Small Bowel Neoplasms	5		Duplication Cysts	2
	Phytobezoar	3		Ileosigmoid Knotting	2
	Internal Herniation	3		Internal Herniation	1
	Foreign Body Ingestion	3		Chilaiditi’s Syndrome	1
	Congenital webs	3		Hirschsprung’s Disease	1
	Ileo-jejunal Volvulus	2		Toxic Megacolon	1
	Jejunal Diverticulitis	2		Colonic Tuberculosis	1
	Gossypiboma	2		Colonic Actinomycosis	1
	Incarceration	2		Enterolithiasis	1
	GIST	1		Brid Ileus	1
	Abdominal Cocoon Syndrome	1		Liposarcoma	1
	Ascariasis	1		Lipoma	1
	Small Bowel Adenocarcinoma	1		Syphilitic Proctitis	1
	SMA Syndrome	1		Ameboma	1
	Parastomal Hernia	1		Dyssynergic Defecation	1
Perforation (n=21)	Trauma	3	Abdominal Pain (n=17)	Appendix Benign Tumors	4
	Foreign Body Ingestion	3		Epiploic Appendagitis.	2
	Medical Instrumentations	3		Internal Herniation	1
	Meckel’s Diverticulum	2		Crohn’s Disease	1
	Jejunal Diverticulum	2		Cecal Duplication Cyst	1
	Crohn’s Disease	2		Inflammatory	1
	GIST	1		Aspergillus Appendicitis	1
	Everolimus Usage	1		Appendix Duplication	1
	Metastatic Melanoma	1		Appendiceal Malignant Tumors	1
	Systemic Lupus	1		COVID-19 Related Colitis	1
	Ileal Diverticulum	1		Schwannoma	1
	Spontaneous Retroperitoneal Perforation	1		Appendix Endometriosis	1
				Cecal Diverticulitis	1
Abdominal Pain (n=5)	Meckel’s Diverticulum	2	Perforation (n=15)	Foreign Body	5
	Ileal Hydatid Cyst	1		Blunt Trauma	3
	Jejunal Diverticulitis	1		Colonic Malignancies	2
	Leiomyosarcoma	1		Collagenous Colitis	1
		Ulcerative Colitis		1	
		Internal Herniation		1	
		COVID-19 Related Colitis		1	
		Chronic Idiopathic Megacolon	1		
Bleeding (n=5)	GIST	3	Fistula (n=9)	Colorectal Malignancies	5
	Metastatic Lung Cancer	1		Colonic Diverticulitis	2
	Schwannoma	1		Anastomosis Leakage	1
		Crohn’s Disease		1	
Fistula (n=2)	Gossypiboma	1	Bleeding (n=3)	GIST	1
	Prolene Mesh Related	1		Inferior Mesenteric Artery (IMA)	1
				Rectal Ischemia	1
Small Bowel Necrosis	Paraduodenal Hernia	1	Ischemia (n=2)	Ischemic Colitis	1
				Stercoral Colitis	1
Total number of cases n=179					

Outcomes

We initially input the following request into the GPT-4 architecture (August 3 Version; OpenAI; <https://chat.openai.com/>): “As healthcare professionals, we plan to utilize you for research purposes. Assuming you are a hypothetical gastrointestinal surgeon, please walk us through the process from differential diagnosis to the most likely disease step by step, based on the patient's information we are about to present”. Following these instructions, the medical history and imaging findings of each patient were entered and the responses from ChatGPT-4 for each case were obtained. Subsequently, both the differential diagnoses and the single final diagnosis were extracted from the answers (an illustrative example is shown in Figure 2). To

mitigate the influence of prior answers on ChatGPT-4's output, we initiated a new ChatGPT-4 session for each case. This procedure was followed for every case during January 2024. ChatGPT-4 output offered several differential diagnoses along with one final diagnosis, which represented the most likely disease. A gastroenterology surgeon with 22 years of experience and a radiologist with 15 years of experience ascertained whether the actual ground truth was consistent with both the differential diagnoses and the final diagnosis generated by ChatGPT-4. In cases of discrepancy, a final decision was reached by consensus (Figure 2).

<p>You As healthcare professionals, we plan to utilize you for research purposes. Assuming you are a hypothetical professional colleague, please walk us through the process from differential diagnosis to the most likely disease step by step, based on the patient's information we are about to present.</p> <p>ChatGPT Certainly! I'm ready to assist you in this hypothetical scenario. Please provide me with the patient's information, including symptoms, medical history, and any other relevant details. Once I have this information, I'll guide you through the process of forming a differential diagnosis and narrowing it down to the most likely disease.</p> <p>A</p>	<p>You HISTORY: 56-year-old male, was informed by general practitioner to the emergency department (ED) of our hospital with progressive abdominal pain over the past 48 h. He was seen after accidental ingestion of fish scales and marrow-rich scraps when eating fish stew 2 days back, and was advised for admission and observation with no other symptoms. The patient admitted to persistent pain in lower abdomen, loss of appetite and nausea which worsened over the past 48 h.</p> <p>On presentation to the ED, the patient was afebrile with stable vital signs and generalized abdominal pain. The laboratory parameters showed increased mild enzyme elevation (lipase 172.5 U, creatinine 78.5 µmol/l) and normal hematological, renal, and hepatic profiles.</p> <p>IMAGING FINDINGS: Oral contrast-enhanced CT scanning promptly demonstrated the presence of free intraperitoneal air and fatty infiltration around the ileocecal anastomosis with extension to the distal sigmoid colon.</p> <p>B</p>
<p>C Differential Diagnosis</p> <p>1...Gastrointestinal Perforation: The presence of free intraperitoneal gas is highly indicative of a perforation in the gastrointestinal tract. This could be related to the accidental ingestion of fish scales/scraps.</p> <p>2...Peritonitis: While the absence of generalized peritonitis signs is present, air leakage and inflammation suggest a localized peritonitis, potentially secondary to micro-perforation and inflammation in sigmoid colon.</p> <p>3...Pancreatitis: Mildly elevated lipase. The ingestion of the bowel could have caused mechanical injury or stress, but without typical radiological findings.</p> <p>4...Diverticulitis/Appendicitis: The specific location of changes around the ileocecal anastomosis raises concern for diverticulitis or appendicitis as differential, although the free intraperitoneal air is not consistent.</p> <p>5...Ischemic Colitis: This could also explain the acute presentation and severe pain, but typically has a different imaging profile.</p>	<p>D Most Likely Diagnosis</p> <p>Given the history of foreign body ingestion (marrow/fish scales), the acute onset of symptoms, and the imaging finding of free intraperitoneal gas and local inflammatory reaction, the most likely diagnosis is gastrointestinal perforation secondary to ingested foreign body. This is supported by the elevated inflammatory markers, which indicate an acute inflammatory process.</p>

Figure 2. Illustration of differential diagnoses and final diagnosis derived from answers

Statistical Analysis

SPSS 25.0 (IBM Corporation, Armonk, New York, United States) program was used in the analysis of variables. The Shapiro-Wilk Francia test was used to evaluate the conformity of the data to normal distribution, while the Levene test was used to evaluate the homogeneity of variance. In comparing quantitative variables according to their diagnostic accuracy, the Independent-Samples T-test was used with Bootstrap results, while the Mann-Whitney U test was used with Monte Carlo results. In comparing diagnostic success according to subgroups of categorical variables, the Ne-Sample Binomial Test was used, while the Pearson Chi-Square and Fisher-Freeman-Halton tests were tested with the Monte Carlo Simulation technique to test the relationship between diagnostic success and other categorical variables. Quantitative variables are expressed in the tables as mean (standard deviation) for those showing normal distribution and as Median (Minimum-Maximum) for those not showing normal distribution, while categorical variables are shown as n (%). Variables were examined at a 95% confidence level, and p-values less than 0.05 were considered significant.

Results

ChatGPT-4 successfully generated differential diagnoses and provided one final diagnosis in all 231 cases. A minimum of 2 and 10 differential diagnoses were generated per case; the median diagnosis was 5 (Q1-Q3:

4-5) and the mean was 4.8 ± 1.2 . A total of 231 cases were included in the study, of which 135 (58.4%) were male and 96 (41.6%) were female. The youngest patient was a 1-year-old boy, while the oldest was a 93-year-old female. The oldest male case was 90 years old. The median age for all cases was 48 (Q1-Q3: 32-65) with a median age for males of 48 (Q1-Q3: 33-65), and for females of 49 (Q1-Q3: 31-67), respectively. There was no statistically significant relationship between age or number of differential diagnoses and gender ($p = 0.687$ and $p = 862$).

Of the 231 cases, 171 (74%) were emergency surgeries and 60 (26%) were elective. Emergency cases ranged from 1 to 93 years in age, while the age range of elective cases was 3 to 84 years. There was no statistical difference in the relationship between the emergency-elective status of cases and their age ($p = 0.845$). Among males, 98 (72.6%) cases were emergency and 37 (27.4%) cases were elective, while in females, 73 (76.0%) cases were emergency and 23 (24.0%) cases were elective. No statistical difference was found between gender and the emergency-elective surgery status of the cases ($p = 0.556$).

The anatomical localization of the cases according to gender and admission type is given in Table 3. Accordingly, the anatomical distribution of the cases is similar according to gender ($p = 0.140$). However, a significant difference was found between the anatomical localizations according to the urgency of the cases ($p < 0.001$).

Table 3: The anatomical localization of the cases according to gender and admission type

	Localization				p
	Esophagus	Stomach	Small intestine	Colorectal	
	n (%)	n (%)	n (%)	n (%)	
Gender					0.140
Male	18 (13.3)	17 (12.6)	45 (33.3)	55 (40.7)	
Female	11 (11.5)	6 (6.3)	45 (46.9)	34 (35.4)	
Surgery status					<0.001
Emergency	11 (6.4)	10 (5.8)	76 (44.4)	74 (43.3)	
Elective	18 (30.0)	13 (21.7)	14 (23.3)	15 (25.0)	
p-value (emergency vs elective)	<0.001	<0.001	0.004	0.012	

Pearson Chi-Square Test (Monte Carlo)

The differential diagnoses were also evaluated, with the ChatGPT-4 program (during both the trained and non-trained periods) providing correct diagnoses in 192 cases (diagnostic accuracy: 83.1%, $p < 0.001$) and incorrect diagnoses in 39 cases (diagnostic accuracy: 16.9%). The

rate of correct differential diagnosis for males was 115 out of 135 (DA: 85.2%, $p < 0.001$), and for females 77 out of 96 (DA: 80.2%, $p < 0.001$). Additionally, ChatGPT-4 accurately detected

differential diagnoses in 142 of 171 emergency cases (DA: 83.1%, $p < 0.001$) and 50 of 60 elective cases (DA: 83.3%, $p < 0.001$). Finally, the correct differential diagnosis rates for esophagus, stomach, small intestine, and colon cases were 82.8% (24 of 29, $p = 0.001$), 78.3% (18 of 23, $p = 0.011$), 84.4% (76 of 90, $p = 0.001$), and 83.1% (74 of 89, $p = 0.001$) respectively (Table 4). Similarly, with respect to the final diagnosis, the ChatGPT-4 program (Trained and Non- trained Period) made a correct diagnosis in 132 out of 231 cases (DA: 57.1%), while making an incorrect diagnosis in 99 cases

($p = 0.035$). ChatGPT-4's correct diagnosis rate for males was 82 out of 135 (DA: 60.7%, $p = 0.016$), and for females was 50 out of 96 (DA: 52.1%, $p = 0.759$). The same software found the correct diagnosis in 96 of the 171 emergency cases (DA: 56.1%, $p = 0.126$), and 36 of the 60 elective cases (DA: 60%, $p = 0.156$). Concerning specific GIS regions, ChatGPT-4 accurately diagnosed 22 out of 29 esophagus cases (DA: 75.9%, $p = 0.009$), 12 out of 23 stomach cases (DA: 52.2%, $p = 0.999$), 50 out of 90 small intestine cases (DA: 55.5%, $p = 0.343$), and 48 out of 89 colon cases (DA: 53.9%, $p = 0.525$) (Table 4).

Table 4: Diagnostic accuracy of all cases (Trained and Non-Trained period of ChatGPT-4)

Trained and Non-trained Period of Chat GPT (n=231)		Differential Diagnosis		P	Final diagnosis		P
		No (n=39)	Yes (n=192)		Incorrect (n=99)	Correct (n=132)	
Diagnosis, n (%)		39 (16.9)	192 (83.1)	<0.001 ^g	99 (42.9)	132 (57.1)	0.035 ^g
Age, med. (min-max)		54 (18-90)	48 (1-93)	0.197 ^u	47 (3-90)	51.5 (1-93)	0.526 ^u
Gender, n (%)				0.374 ^c			0.225 ^c
	Male	20 (14.8)	115 (85.2)	<0.001 ^g	53 (39.3)	82 (60.7)	0.016 ^g
	Female	19 (19.8)	77 (80.2)	<0.001 ^g	46 (47.9)	50 (52.1)	0.759 ^g
Surgery status, n (%)				0.999 ^c			0.651 ^c
	Emergency	29 (17)	142 (83)	<0.001 ^g	75 (43.9)	96 (56.1)	0.126 ^g
	Elective	10 (16.7)	50 (83.3)	<0.001 ^g	24 (40)	36 (60)	0.156 ^g
Location, n (%)				0.870 ^{ff}			0.184 ^c
	Esophagus	5 (17.2)	24 (82.8)	0.001 ^g	7 (24.1)	22 (75.9)	0.009 ^g
	Stomach	5 (21.7)	18 (78.3)	0.011 ^g	11 (47.8)	12 (52.2)	0.999 ^g
	Small intestine	14 (15.6)	76 (84.4)	<0.001 ^g	40 (44.4)	50 (55.6)	0.343 ^g
	Colon	15 (16.9)	74 (83.1)	<0.001 ^g	41 (46.1)	48 (53.9)	0.525 ^g
Number of differential diagnoses, med. (min-max)		4 (2-9)	5 (3-10)	0.002 ^u	5 (2-9)	5 (3-10)	0.350 ^u

One-Sample Binomial Test, Mann-Whitney U Test (Monte Carlo), Fisher-Freeman-Halton Test (Monte Carlo), Pearson Chi-Square Test (Monte Carlo), med.: Median, SD: Standard Deviation. Discrete and continuous variables that did not follow a normal distribution are presented as med. (min-max), while those that followed a normal distribution are presented as mean (SD).

Considering that the ChatGPT-4 learning practice was completed in April 2023, the cases were further divided into before and after April 2023. 164 cases (Group A- as trained period data) belonged to the period before April 2023 and 67 cases (Group B- as non-trained period data) belonged to that after April 2023. These groups were compared among themselves. There were 91 males and 73 females in Group A, and 44 males and 23 females in Group B ($p = 0.154$). In Group A, 120 cases were urgent and 44 cases were elective, in Group B, 51 cases were urgent and 16 cases were elective ($p = 0.643$). In addition, Groups A and B were similar in the frequency

of anatomical localization of cases, the diagnostic accuracy rate of final diagnosis, and the probability of differential diagnoses to predict final diagnosis (Table 5). ChatGPT-4 made correct predictions in 97 cases in Group A (diagnostic accuracy: 59.1%) and in 35 cases in Group B (diagnostic accuracy: 52.2%) ($p = 0.904$). Additionally, the mean age and the number of differential diagnoses were statistically similar between Groups A and B ($p = 0.877$ and $p = 0.556$). An evaluation was made for Group A (Trained period) in terms of differential diagnosis, and the ChatGPT-4 program predicted correct diagnoses in 136 cases (DA: 82.9%, $p <$

0.001) and incorrect diagnoses in 28 cases (DA: 17.1%). In Group A, correct differential diagnose was made in 76 of 91 male patients (DA: 83.5%, $p < 0.001$) and 60 of 73 female patients (DA: 80.2%, $p < 0.001$). For the same group, ChatGPT-4 provided an accurate differential diagnosis in 99 of 120 emergency cases (DA: 82.5%, $p < 0.001$) and 37 of 44 elective cases (DA: 84.1%, $p < 0.001$). Finally, ChatGPT-4 made the correct differential diagnosis in 18 of 23 esophagus cases (DA: 78.2%, $p < 0.001$), 10 of 13 stomach cases (DA: 76.9%, $p < 0.052$), 57 of 67 small intestine cases (DA: 85.1%, $p < 0.001$), and 51 of 61 colon cases (DA: 83.6%, $p < 0.001$) (Table 6). For

cases before April 2023, the ChatGPT-4 program made correct final diagnoses in 97 cases (DA: 59.1%, $p = 0.024$). The DA for the final diagnosis was significant for males and for the esophagus ($p = 0.012$ and $p = 0.035$, respectively). For Group B (Non-Trained period) the rate of DA and significance levels are given in Table 7. The DA for the differential diagnosis of ChatGPT-4 was 83.6%, while for the final diagnosis it was 52.2% ($p < 0.001$ and $p = 0.807$, respectively). The DA for the differential diagnosis of esophageal disease was 100%, whereas the DA for the final diagnosis was 83.3% ($p = 0.102$).

Table 5: Comparison of findings for Trained and Non-Trained period of ChatGPT-4

	ChatGPT4 period			p
	Trained (Group A)		Non-Trained (Group B)	
	n (%)		n (%)	
Gender				0.154
Male	91 (55.5)		44 (65.7)	
Female	73 (44.5)		23 (34.3)	
Surgery status				0.643
Emergency	120 (73.2)		51 (76.1)	
Elective	44 (26.8)		16 (23.9)	
Localization				0.251
Esophagus	23 (14.0)		6 (9.0)	
Stomach	13 (7.9)		10 (14.9)	
Small intestine	67 (40.9)		23 (34.3)	
Colorectal	61 (37.2)		28 (41.8)	
FD				0.904
Incorrect	67 (40.9)		32 (47.8)	
Correct	97 (59.1)		35 (52.2)	
FD in DDs.				0.336
Yes	28 (17.1)		11 (16.4)	
No	136 (82.9)		56 (83.6)	

Pearson Chi-Square Test (Monte Carlo)

Table 6: Diagnostic accuracy of trained period of ChatGPT-4

Trained Period of Chat GPT (n=164)	Differential diagnosis		p	Final diagnosis		p
	No (n=28)	Yes (n=136)		Incorrect (n=67)	Correct (n=97)	
Diagnosis, n (%)	28 (17.1)	136 (82.9)	<0.001^g	67 (40.9)	97 (59.1)	0.024^g
Age, mean(SD)	54.8 (18.8)	47.4 (21.8)	0.063[†]	47.7 (20.7)	49.3 (22.1)	0.623[†]
Gender, n (%)			0.837 ^c			0.203 ^c
Male	15 (16.5)	76 (83.5)	<0.001 ^g	33 (36.3)	58 (63.7)	0.012 ^g
Female	13 (17.8)	60 (82.2)	<0.001 ^g	34 (46.6)	39 (53.4)	0.640 ^g
Surgery status, n (%)			0.823 ^c			0.999 ^c
Emergency	21 (17.5)	99 (82.5)	<0.001 ^g	49 (40.8)	71 (59.2)	0.055 ^g
Elective	7 (15.9)	37 (84.1)	<0.001 ^g	18 (40.9)	26 (59.1)	0.291 ^g
Location, n (%)			0.753 ^{ff}			0.370 ^c
Esophagus	5 (21.7)	18 (78.3)	0.011 ^g	6 (26.1)	17 (73.9)	0.035 ^g
Stomach	3 (23.1)	10 (76.9)	0.052 ^g	7 (53.8)	6 (46.2)	0.999 ^g
Small intestine	10 (14.9)	57 (85.1)	<0.001 ^g	29 (43.3)	38 (56.7)	0.328 ^g
Colon	10 (16.4)	51 (83.6)	<0.001 ^g	25 (41)	36 (59)	0.200 ^g
Number of differential diagnoses, med. (min-max)	4 (3-9)	5 (3-10)	0.046^u	5 (3-9)	5 (3-10)	0.796^u

^g One-Sample Binomial Test, Mann-Whitney U Test (Monte Carlo), Fisher-Freeman-Halton Test (Monte Carlo), Pearson Chi-Square Test (Monte Carlo), Independent t-Test (Bootstrap), med.: Median, SD: Standard Deviation. Discrete and continuous variables that did not follow a normal distribution are presented as med. (min-max), while those that followed a normal distribution are presented as mean (SD).

Table 7: Diagnostic accuracy of non-trained period of ChatGPT-4

Non-trained Period of Chat GPT (n=67)	Differential diagnosis		P	Final diagnosis		P
	No (n=11)	Yes (n=56)		Incorrect (n=32)	Correct (n=35)	
Diagnosis, n (%)	11 (16.4)	56 (83.6)	<0.001^g	32 (47.8)	35 (52.2)	0.807^g
Age, mean(SD) or med. (min-max)	42 (28-90)	50 (1-85)	0.914^u	48 (22.8)	50.2 (24.2)	0.688^t
Gender, n (%)			0.167^f			0.618^c
Male	5 (11.4)	39 (88.6)	<0.001 ^g	20 (45.5)	24 (54.5)	0.651 ^g
Female	6 (26.1)	17 (73.9)	0.035 ^g	12 (52.2)	11 (47.8)	0.999 ^g
Surgery status, n (%)			0.716^f			0.400^c
Emergency	8 (15.7)	43 (84.3)	<0.001 ^g	26 (51)	25 (49)	0.999 ^g
Elective	3 (18.8)	13 (81.3)	0.021 ^g	6 (37.5)	10 (62.5)	0.454 ^g
Location, n (%)			0.819^{ff}			0.326^{ff}
Esophagus	0 (0)	6 (100)	0	1 (16.7)	5 (83.3)	0.102 ^g
Stomach	2 (20)	8 (80)	0.109 ^g	4 (40)	6 (60)	0.058 ^g
Small intestine	4 (17.4)	19 (82.6)	0.003 ^g	11 (47.8)	12 (52.2)	0.999 ^g
Colon	5 (17.9)	23 (82.1)	0.001 ^g	16 (57.1)	12 (42.9)	0.571 ^g
Number of differential diagnoses, med. (min-max)	4 (2-5)	5 (3-9)	0.011^u	5 (2-9)	4 (3-7)	0.189^u

^g One-Sample Binomial Test, Mann-Whitney U Test (Monte Carlo), Fisher-Freeman-Halton Test (Monte Carlo), Pearson Chi-Square Test (Monte Carlo), Independent t-Test (Bootstrap), med.: Median, SD: Standard Deviation, : absolute accuracy. Discrete and continuous variables that did not follow a normal distribution are presented as med. (min-max), while those that followed a normal distribution are presented as mean (SD).

Discussion

Recent advances in artificial intelligence (AI), particularly through large language models (LLMs), are reshaping the future of clinical medicine. ChatGPT-4, as a generative AI model, has been increasingly studied across different medical specialties; however, its performance in gastrointestinal surgery (GIS) remains largely unexplored [2, 13, 26-33]. Our study is among the first to assess ChatGPT-4's diagnostic accuracy (DA) using real-world case reports from The International Journal of Surgery Case Reports, thereby moving beyond hypothetical scenarios or single-disease models explored in earlier studies.

The main findings demonstrated a final DA of 57.1% (132/231 cases) and a differential diagnosis accuracy exceeding 80%. These results suggest that ChatGPT-4, while not a replacement for clinical judgment, can provide meaningful support by broadening the diagnostic perspective, especially in complex or rare cases. For surgeons faced with time-sensitive emergencies, having access to an AI tool capable of suggesting plausible

differential diagnoses may reduce the risk of oversight and improve early decision-making.

Surgical sciences are known for their intense workload and associated diagnostic error risk [34-36]. Under these circumstances, ChatGPT-4 can serve as a valuable adjunct, offering quick access to a broad knowledge base and improving efficiency in both emergency and elective cases. Importantly, this potential support may be particularly beneficial in settings with limited surgical manpower, such as low-income regions or disaster zones [37]. While its role will likely remain complementary, ChatGPT-4 may enhance the allocation of surgeon time, reduce fatigue, and contribute to patient safety.

Beyond diagnostic accuracy, the question of how ChatGPT-4 can be integrated into clinical workflows is crucial. In emergency settings, where surgeons face severe time constraints and limited access to subspecialty consultation, ChatGPT-4 could rapidly generate differential diagnoses, highlight red-flag features, and suggest targeted investigations. For example, in patients with acute abdomen or suspected

perforation, the model may support triage decisions, guide early stabilization steps, and help prioritize life-threatening conditions. Such applications may be particularly valuable for junior physicians or in facilities lacking round-the-clock surgical coverage, ultimately optimizing decision-making during critical minutes.

In resource-limited environments, ChatGPT-4's impact may be even more pronounced. Regions with shortages of trained surgeons, radiologists, or gastroenterologists could use ChatGPT-4 to synthesize available clinical and radiological data into structured diagnostic outputs. By assisting less experienced clinicians, the model could reduce diagnostic uncertainty, standardize care pathways, and improve patient safety. Moreover, in disaster or conflict zones, where both expertise and infrastructure are constrained, ChatGPT-4 could function as an accessible, low-cost adjunct to support frontline providers until expert consultation becomes available. For sustainable integration, however, attention must be given to interoperability with electronic health record (EHR) systems, adaptation to local languages, and offline functionality where internet connectivity is unreliable.

When comparing our findings with existing literature, several important parallels and distinctions emerge. Wei et al. [2], through a systematic review and meta-analysis, reported an integrated accuracy of 56% for ChatGPT's responses to medical queries. This closely aligns with our final DA of 57.1%, suggesting that ChatGPT-4 demonstrates a consistent level of performance across specialties. Similarly, Horiuchi et al. [20] reported 50% accuracy in neuroradiology, underscoring that while ChatGPT can provide useful insights, its accuracy generally remains below that of specialist physicians. Our study extends this evidence into GIS, confirming comparable performance in a surgical subspecialty.

Ueda et al. [21] demonstrated that ChatGPT's accuracy varied substantially depending on the type of input provided (22% with history alone vs. 61% with combined history and imaging). In our study, although full radiologic images were not directly used, reliance on textual case descriptions parallels Ueda's "history-only" condition, which may partially explain the moderate accuracy observed. This highlights the critical role of structured and comprehensive inputs for optimizing ChatGPT's performance.

Lechien et al. [31], who tested ChatGPT in otolaryngology, showed high performance in suggesting primary diagnoses (90%) but reduced accuracy in management planning. This resonates with our finding that differential diagnoses were more reliable (>80%) than final diagnoses (57.1%). Both studies emphasize that ChatGPT's current strength lies more in broadening diagnostic perspectives than in providing definitive diagnoses or treatment recommendations.

In contrast, convolutional neural network (CNN)-based models have achieved higher performance levels in imaging tasks: Swager et al. [38] in Barrett's esophagus (AUC 0.95), Zhang et al. [12] in esophageal lesions (AUC up to 0.907), Kubota et al. [39] in gastric cancer staging (accuracy 64.7%), and Wang et al. [40] in polyp detection (AUC 0.984). Compared with these, ChatGPT-4's text-based diagnostic accuracy is modest. However, its unique strength is handling unstructured narratives and providing conversational, case-wide reasoning, making it complementary to image-based AI systems rather than directly competitive.

Several factors likely contributed to misdiagnoses. First, the case mix consisted largely of rare or atypical etiologies, since the source journal favors unusual educational cases. This increased the chance of ChatGPT-4 defaulting to common conditions when textual descriptors were limited, thereby lowering final diagnosis accuracy compared to differential diagnosis performance. Second, anatomical site effects were evident: esophageal cases contained more distinctive clinical cues, while small intestine and colon cases presented overlapping syndromes such as obstruction or inflammatory disease, which were harder to distinguish based on text alone. Gender-related differences were also noted (higher accuracy in males), though these may reflect sample size variation or subtle differences in case framing rather than inherent bias. Finally, input style and completeness strongly influenced outcomes: structured, detailed case descriptions yielded more accurate outputs, whereas unstructured narratives or missing negative findings (e.g., absence of peritoneal signs) reduced discriminative power. These observations suggest that structured prompts and integration with image-based AI could reduce error rates and enhance reliability in future applications.

Despite promising findings, some limitations must be acknowledged. ChatGPT-4 was not evaluated with direct imaging inputs, and its outputs were compared against case reports that may have been curated to emphasize specific diagnoses, potentially inflating accuracy. Furthermore, while differential diagnoses were reasonably strong, final diagnosis accuracy was lower, indicating that human oversight remains essential. Finally, the lack of transparency in ChatGPT-4's training data and update cycles raises questions about reproducibility and long-term reliability.

Future studies should validate these findings in broader and more routine GIS populations, with larger datasets and real-time clinical applications. Combining ChatGPT-4 with imaging-based AI algorithms may yield greater diagnostic precision and strengthen decision-support systems. Ultimately, while AI cannot substitute clinical expertise, surgeons who embrace such tools may gain a significant advantage, both in patient safety and efficiency.

Conclusions

This study provides one of the first systematic evaluations of ChatGPT-4's diagnostic performance in gastrointestinal surgery (GIS) using real-world patient cases. ChatGPT-4 demonstrated high diagnostic accuracy for differential diagnoses (>80%), but a lower accuracy for final diagnoses (57.1%). These findings underscore its value as a supportive tool, helping to broaden diagnostic perspectives and alert clinicians to less obvious possibilities, while highlighting the need for cautious interpretation when narrowing to a single final diagnosis.

Importantly, ChatGPT-4 should not be regarded as a replacement for clinical judgment. Its role is complementary—supporting surgeons in time-sensitive decision-making, reducing fatigue under heavy workloads, and extending diagnostic resources in low-resource or high-demand settings. As integration into clinical practice evolves, real-time testing, larger datasets, and multimodal approaches combining text with image-based AI will be essential to refine its capabilities. Surgeons who learn to incorporate ChatGPT-4 effectively into their workflow may gain a meaningful advantage in efficiency, safety, and patient outcomes.

Funding: None declared.

Competing interests: The author declares no competing interests.

Acknowledgements: The manuscript was developed with the assistance of Chat Generative Pre-trained Transformer (ChatGPT) based on the GPT-4 architecture (August 3 Version; OpenAI; [<https://chat.openai.com/>])(<https://chat.openai.com/>)), with all outputs verified by the author. Special thanks are extended to Emrah Karatay (Radiology Specialist) for evaluating the images.

References

1. Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI models: a preliminary review. *Futur Internet*. 2023;15(6):1-15.
2. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform*. 2024;151:38462064.
3. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877-901.
4. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: early experiments with GPT-4 [Internet]. arXiv; 2023 [cited 2025 Jan 10]. Available from: <http://arxiv.org/abs/2303.12712>(<http://arxiv.org/abs/2303.12712>)
5. Eloundou T, Manning S, Mishkin P, Rock D. GPTs are GPTs: an early look at the labor market impact potential of large language models [Internet]. arXiv; 2023 [cited 2025 Jan 10]. Available from: <http://arxiv.org/abs/2303.10130>(<http://arxiv.org/abs/2303.10130>)
6. Suchman K, Garg S, Trindade AJ. Chat generative pretrained transformer fails the multiple-choice American College of Gastroenterology self-assessment test. *Am J Gastroenterol*. 2023;118(12):2280-2.
7. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology*. 2023;307(4):2-5.
8. Gertz RJ, Bunck AC, Lennartz S, Dratsch T, Iuga AI, Maintz D, et al. GPT-4 for automated determination of radiologic study and protocol based on radiology request forms: a feasibility study. *Radiology*. 2023;307(5):4-6.
9. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retin*. 2023;7(10):862-8.
10. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. *J Med Internet Res*. 2023;25:2023-4.
11. Weese JS, Scherk M, Hostutler RA, Armstrong J, Heseltine J, Evason M, et al. Applications of artificial intelligence for the diagnosis of gastrointestinal diseases. In: *A color handbook of infectious diseases of the dog and cat*. London: CRC Press; 2019. p. 49-96.
12. Zhang M, Zhu C, Wang Y, Kong Z, Hua Y, Zhang W, et al. Differential diagnosis for esophageal protruded lesions using a deep convolution neural network in endoscopic images. *Gastrointest Endosc*. 2021;93(6):1261-72.e2.
13. Datt M, Sharma H, Aggarwal N, Sharma S. Role of ChatGPT-4 for medical researchers. *Ann Biomed Eng*. 2024;52(6):1534-6.
14. Tang Y, Zhong X, Su X, Hu X, Li X. Is ChatGPT-4 a qualified surgical oncologist? *Int J Surg*. 2022;105:2846-8.
15. Abadir AP, Ali MF, Karnes W, Samarasekera JB. Artificial intelligence in gastrointestinal endoscopy. *Clin Endosc*. 2020;53(2):132-41.
16. El Hajjar A, Rey JF. Artificial intelligence in gastrointestinal endoscopy: general overview. *Chin Med J (Engl)*. 2020;133(3):326-34.
17. Messmann H, Bisschops R, Antonelli G, Libanio D, Sinouque P, Abdelrahim M, et al. Expected value of artificial intelligence in gastrointestinal endoscopy: European Society of Gastrointestinal Endoscopy (ESGE) position statement. *Endoscopy*. 2022;54(12):1211-31.
18. Correia FP, Lourenço LC. Artificial intelligence application in diagnostic gastrointestinal endoscopy: deus ex machina? *World J Gastroenterol*. 2021;27(32):5351-61.
19. Kuo CY, Chiu HM. Application of artificial intelligence in gastroenterology: potential role in clinical practice. *J Gastroenterol Hepatol*. 2021;36(2):267-72.
20. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S, et al. Accuracy of ChatGPT-generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. *Neuroradiology*. 2024;66(1):73-9.
21. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the Diagnosis Please quizzes. *Radiology*. 2023;308(3):1-2.
22. Khanna RK, Ducloyer JB, Hage A, Rezkallah A, Durbant E, Bigoteau M, et al. Evaluating the potential of ChatGPT-4 in ophthalmology: the good, the bad and the ugly. *J Fr Ophtalmol*. 2023;46(7):697-705.
23. Delsoz M, Madadi Y, Raja H, Munir WM, Tamm B, Mehravaran S, et al. Performance of ChatGPT in diagnosis of corneal eye diseases. *Cornea*. 2024;43(5):664-70.
24. Kerbage A, Kassab J, El Dahdah J, Burke CA, Achkar JP, Rouphael C. Accuracy of ChatGPT in common gastrointestinal diseases: impact for patients and providers. *Clin Gastroenterol Hepatol*. 2024;22(6):1323-5.e3.
25. Cheng LK, O'Grady G, Du P, Egbuji JU, Windsor JA, Pullan AJ. Gastrointestinal system. *Wiley Interdiscip Rev Syst Biol Med*. 2010;2(1):65-79.
26. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-40.
27. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Kartowicz J. Beyond ChatGPT: what does GPT-4 add to healthcare? The dawn of a new era. *Cardiol J*. 2023;30(6):1018-25.
28. Gutiérrez-Cirlos C, Bermúdez-González JL, Carrillo-Pérez DL, Hidrogo-Montemayor I, Martínez-González A, Carrillo-Esper R, et al. Medicine and the metaverse: current applications and future. *Gac Med Mex*. 2023;159(4):376-92.
29. Jiao C, Edupuganti NR, Patel PA, Bui T, Sheth V. Evaluating the artificial intelligence performance growth in ophthalmic knowledge. *Cureus*. 2023;15(9):e37868.
30. Leybold T, Schäfer B, Boos A, Beier JP. Can AI think like a plastic surgeon? Evaluating GPT-4's clinical judgment in reconstructive procedures of the upper extremity. *Plast Reconstr Surg Glob Open*. 2023;11(12):e5471.
31. Lechien JR, Georgescu BM, Hans S, Chiesa-Estomba CM. ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. *Eur Arch Otorhinolaryngol*. 2024;281(1):319-33.
32. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg*. 2018;268(1):70-6.
33. Beaulieu-Jones BR, Berrigan MT, Shah S, Marwaha JS, Lai SL, Brat GA. Evaluating capabilities of large language models: performance of GPT-4 on surgical knowledge assessments. *Surgery*. 2024;175(4):936-42.
34. Jalali M, Esmaeili R, Habibi E, Alizadeh M, Karimi A. Mental workload profile and its relationship with presenteeism, absenteeism and job performance among surgeons: the mediating role of occupational fatigue. *Heliyon*. 2023;9(9):e20234.
35. Blum J, Menta AK, Zhao X, Yang VB, Gouda MA, Subbiah V. Pearls and pitfalls of ChatGPT in medical oncology. *Trends Cancer*. 2023;9(10):788-90.
36. Liu JH, Etzioni DA, O'Connell JB, Maggard MA, Ko CY. The increasing workload of general surgery. *Arch Surg*. 2004;139(4):423-8.
37. Salam A, Wireko AA, Jiffry R, Ng JC, Patel H, Zahid MJ, et al. The impact of natural disasters on healthcare and surgical services in low- and middle-income countries. *Ann Med Surg*. 2023;85:3774-7.

38. Swager AF, van der Sommen F, Klomp SR, Zinger S, Meijer SL, Schoon EJ, et al. Computer-aided detection of early Barrett's neoplasia using volumetric laser endomicroscopy. *Gastrointest Endosc.* 2017;86(5):839-46.
39. Kubota K, Kuroda J, Yoshida M, Ohta K, Kitajima M. Medical image analysis: computer-aided diagnosis of gastric cancer invasion on endoscopic images. *Surg Endosc.* 2012;26(5):1485-9.
40. Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat Biomed Eng.* 2018;2(10):741-8.