Thieme

Using a customized GPT to provide guideline-based recommendations for management of pancreatic cystic lesions

Authors
Yuri Gorelik[1], Itai Ghersin[1], Tarek Arraf[1], Offir Ben-Ishay[2], Amir Klein‡[1], Iyad Khamaysi‡[3]

Institutions
1 Department of Internal Medicine D, Gastroenterology, Rambam Health Care Campus, Haifa, Israel
2 Surgery, Rambam Health Care Campus, Haifa, Israel
3 Gastroenterology, Rambam MC, Kfar Kana, Israel

Keywords
Pancreas, Endoscopic ultrasonography, Fine-needle aspiration/biopsy, Pancreatobiliary (ERCP/PTCD), MRCP topics

received 2.2.2024
accepted after revision 15.3.2024
accepted manuscript online 18.3.2024

Bibliography
Endosc Int Open 2024; 12: E600–E603
DOI 10.1055/a-2289-9334
ISSN 2364-3722

Georg Thieme Verlag KG, Rüdigerstraße 14,
70469 Stuttgart, Germany

Corresponding author
Yuri Gorelik, MD. MPH, Rambam Health Care Campus – Department of Internal Medicine D, Gastroenterology, Haaliya Hashnia Haifa 31999, Israel
yurigorelik@gmail.com

ABSTRACT

Background and study aims Rising prevalence of pancreatic cysts and inconsistent management guidelines necessitate innovative approaches. New features of large language models (LLMs), namely custom GPT creation, provided by ChatGPT can be utilized to integrate multiple guidelines and settle inconsistencies.

Methods A custom GPT was developed to provide guideline-based management advice for pancreatic cysts. Sixty clinical scenarios were evaluated by both the custom GPT and gastroenterology experts. A consensus was reached between experts and review of guidelines and the accuracy of recommendations provided by the custom GPT was evaluated and compared with experts.

Results The custom GPT aligned with expert recommendations in 87% of scenarios. Initial expert recommendations were correct in 97% and 87% of cases, respectively. No significant difference was observed between the accuracy of custom GPT and the experts. Agreement analysis using Cohen's and Fleiss' Kappa coefficients indicated consistency among experts and the custom GPT.

Conclusions This proof-of-concept study shows the custom GPT's potential to provide accurate, guideline-based recommendations for pancreatic cyst management, comparable to expert opinions. The study highlights the role of advanced features of LLMs in enhancing clinical decision-making in fields with significant practice variability.

‡ These authors contributed equally.

Introduction

The prevalence of pancreatic cysts is steadily increasing, which is presumed to result partially from the proliferation of imaging studies [1]. The approach to surveillance and management of these cysts is controversial, with notable inconsistencies among society guidelines and management practices. There is also evidence of wide variation in surveillance [2, 3, 4, 5, 6, 7].

With the advent of advanced deep learning algorithms, large language models (LLMs) like ChatGPT have become widely used in various fields, including medicine [8]. The application of ChatGPT in medical disciplines is rapidly expanding [9]. Its potential as a resource for gastroenterologists is also becoming evident; for instance, we have recently showcased its effectiveness in improving management of patients after colonoscopy with polypectomy [10]. Recently, introduction of new features

has enabled the creation of custom GPTs - a tailored version of ChatGPT with specialized knowledge and directives.

We aimed to assess the effectiveness of a custom GPT in providing integrated guideline-based recommendations for managing pancreatic cysts.

Methods

This was a proof-of-concept study in which a customized GPT was configured and tuned. We then created and tested the recommendations the GPT provided for management of various clinical scenarios involving pancreatic cystic lesions.

GPT creation

We developed a GPT named "Pancreatic Cyst Recommendations," specifically programmed to offer guideline-based advice for managing pancreatic cysts. Custom GPT creation and update are accessible and intuitive tasks that do not require programming skills and the generated GPT can be publicly shared. The GPT was instructed to prioritize the international consensus guidelines [3], and the guidelines from the American Gastroenterological Association [2], for each case. In instances of significant discrepancies, it was programmed to consider the American College of Gastroenterology guidelines and the joint European guidelines [4, 5]. This methodology mirrors our routine clinical approach. To further emulate real-world practice and ensure consistency, we instructed the GPT to provide one of the following recommendations for each scenario: "no further surveillance," "periodic surveillance with imaging," "endoscopic ultrasound (EUS) with fine needle aspiration (FNA) of the cyst," or "referral for surgical evaluation." The relevant guidelines were uploaded to the GPT as PDF files. The system's configuration was fine-tuned based on responses to a set of preliminary scenarios, which were excluded from the final analysis.

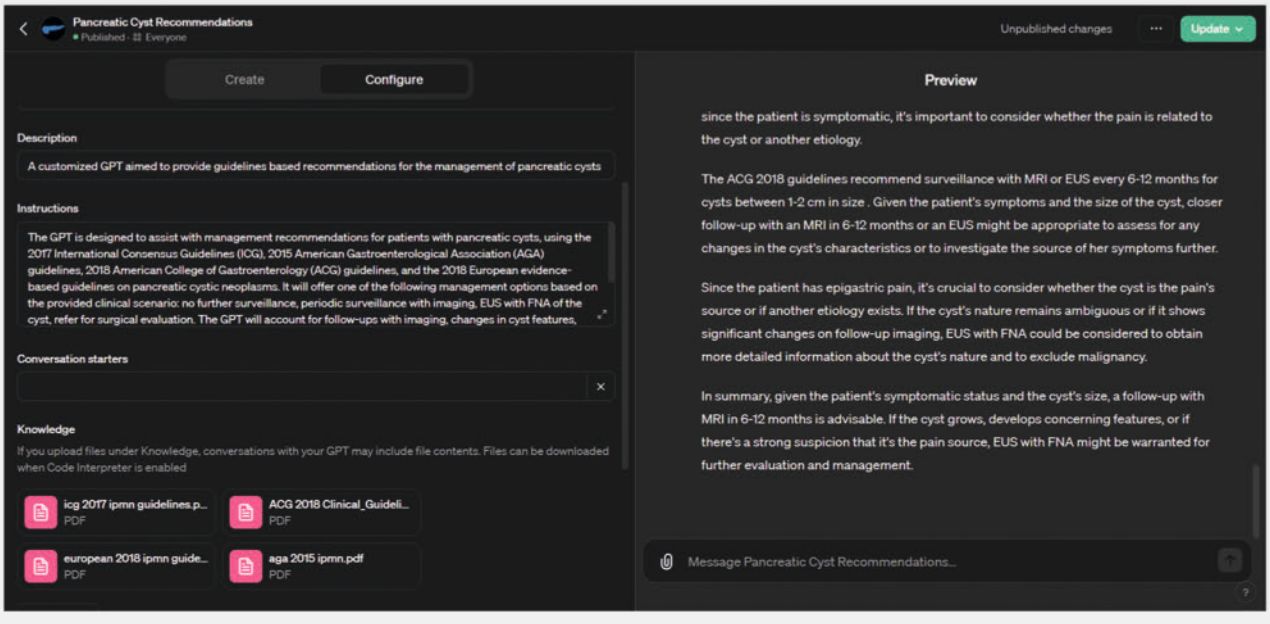
The GPT can be evaluated using the following link: https://chat.openai.com/g/g-xXZ94NXwW-pancreatic-cyst-recommendations (▶ **Fig. 1**)

Clinical scenarios

Sixty clinical scenarios were devised for this study: 30 by a team of three gastroenterologists (TA, IK, and YG) and another 30 through a separate ChatGPT session. To generate these scenarios, we requested detailed clinical cases encompassing patient history, physical examination findings, relevant laboratory and imaging study results including follow-ups, and, where applicable, fluid analysis results. ChatGPT's proficiency in creating suitable clinical vignettes had been previously assessed [11]. We conducted a thorough review of each vignette generated by ChatGPT to ensure its appropriateness.

Evaluation of recommendations

Each scenario was presented to two gastroenterologists (pancreato-biliary specialists), a hepato-biliary surgeon and the customized GPT agent. These physicians provided recommendations from one of the four predefined options (as detailed in the GPT creation section). In addition, three gastroenterologists assessed the scenarios against each of the four guidelines. The determination of the correct answer for each scenario was conducted as follows: 1) If both senior gastroenterologists provided the same recommendation, and this recommendation aligned with the guidelines, it was deemed correct; 2) If the gastroenterologists offered different recommendations, or different from the surgeon, but both were consistent with the



▶ **Fig. 1** A screenshot of the custom GPT configuration (left panel) with an example of a scenario prompt and answer (right panel).

guidelines, both answers were considered correct; and 3) A consensus meeting was held for cases where the physicians provided differing recommendations, and one or both were not in accordance with the guidelines. A mutually agreed upon correct answer was established for each such scenario.

The outcomes were the accuracy of the GPT's recommendations and their comparison with the accuracy of expert recommendations before consensus. Additional outcomes were concordance levels between the custom GPT and expert responses and concordance within experts.

Statistical analysis

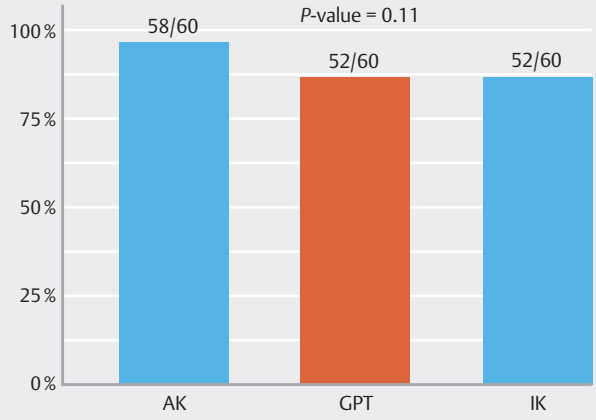
The proportion of clinical scenarios where the GPT's, and the experts' recommendations matched the correct recommendations (following consensus) were calculated with confidence intervals (Cis) calculated using the Wilson score method. The rates were compared using the Kruskal-Wallis test.

To assess the agreement level with each gastroenterologist, and to calculate the agreement between guidelines, we calculated Cohen's Kappa, and Fleiss' Kappa coefficients, respectively. We performed subgroup analyses based on the origin of the clinical scenarios (developed by gastroenterologists vs. generated by ChatGPT). All statistical analyses were performed using R (version 4.2.0, R Foundation for Statistical Computing, Vienna, Austria). $P < 0.05$ was considered statistically significant.

Results

Initial physician responses were concordant in 65% of scenarios (39/60), with a lower concordance between gastroenterologists and the hepatobiliary surgeon (mean of 53% [32/60]). The Cohen's Kappa coefficients for the agreement between IK and AK was 0.68 (95% CI, 0.51–0.85). Cohen's Kappa coefficients for the agreement between the GPT agent and the gastroenterologists was 0.65 (95% CI, 0.49–0.81) for concordance with AK, and 0.61 (95% CI 0.44–0.78) for concordance with IK.

A consensus between experts was then reached in all the disconcordant cases, thus creating a final set of correct recommendation to which the GPT was compared. Correct answers rate and comparisons are presented in ▶ **Fig. 2**. The GPT agent provided correct recommendations in 87% (52 of 60, 95% CI 76%-93%). The initial expert recommendations, pre-consensus, were correct in 97% (58 of 60, 95% CI 76%-99%) and 87% (52 of 60, 95% CI 76%-93%), respectively. With the limitation of this small sample size, no significant differences were observed between the custom GPT and experts ($P = 0.11$).



▶ **Fig. 2** Bar chart showing accuracy rates of the recommendations provided by the custom GPT and the experts.

Discussion

In this proof of concept, a custom-trained GPT agent demonstrated the capability to provide guideline-based recommendations for managing pancreatic cysts with high accuracy, correctly addressing 52 of 60 scenarios, a rate that did not substantially differ from the rate achieved by recommendations provided by experts. The variability seen in this study between physicians and different guidelines further underscores the significant variability in the practice and management of pancreatic cysts. These discrepancies between guidelines and low adherence rates among physicians to these guidelines are well noted limitations of this field [7, 12]. For these reasons, we believe that an integrated LLM-based solution is potentially very helpful in this clinical field.

A notable advantage of the custom GPT is its adaptability, which allows for tailoring of the recommendations (i. e. more conservative or more assertive, depending on patient-specific risk factors, local epidemiology and resources). Custom GPTs can be easily reconfigured to use updated guidelines and practices as they are introduced. For example, it was recently suggested that discontinuing surveillance of branch duct intraductal papillary mucinous neoplasms lacking worrisome features or high-risk stigmata is justified [13].

The primary limitation of this study is the inconsistency between physician opinions and existing guidelines, which complicates the establishment of a definitive standard against which to compare the custom GPT. This difficulty echoes the inherent dilemmas in this field. Our methodology, which involves juxtaposing guidelines with consensus meetings, integrates expert physician insights with varied guidelines, providing a robust standard for assessing the custom GPT's accuracy. Furthermore, to the best of our knowledge, this is the practice in most academic centers worldwide.

Another constraint is that the custom GPT was evaluated using simulated scenarios instead of actual clinical cases collected either retrospectively or prospectively. In creating these scenarios, we engaged multiple gastroenterologists and utilized ChatGPT to ensure a broad spectrum of cases and features. Future research should focus on applying the custom GPT to real-world clinical data, both retrospective and prospective evaluations.

Conclusions

In conclusion, our findings highlight the variability in current management of pancreatic cysts and demonstrate that a custom GPT can provide highly accurate recommendations as compared to the integration of medical experts and guidelines.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Schweber AB, Agarunov E, Brooks C et al. Prevalence, incidence, and risk of progression of asymptomatic pancreatic cysts in large sample real-world data. Pancreas 2021; 50: 1287–1292 doi:10.1097/MPA.0000000000001918

[2] Vege SS, Ziring B, Jain R et al. American gastroenterological association institute guideline on the diagnosis and management of asymptomatic neoplastic pancreatic cysts. Gastroenterology 2015; 148: 819–822 doi:10.1053/j.gastro.2015.01.015

[3] Tanaka M, Fernández-del Castillo C, Kamisawa T et al. Revisions of international consensus Fukuoka guidelines for the management of IPMN of the pancreas. Pancreatology 2017; 17: 738–753

[4] Del Chiaro M, Besselink MG, Scholten L et al. European evidence-based guidelines on pancreatic cystic neoplasms. Gut 2018; 67: 789–804

[5] Elta GH, Enestvedt BK, Sauer BG et al. ACG Clinical Guideline: Diagnosis and Management of Pancreatic Cysts. Am J Gastroenterol 2018; 113: 464–479 doi:10.1038/ajg.2018.14

[6] van Huijgevoort NCM, del Chiaro M, Wolfgang CL et al. Diagnosis and management of pancreatic cystic neoplasms: current evidence and guidelines. Nat Rev Gastroenterol Hepatol 2019 1611 2019; 16: 676–689 doi:10.1038/s41575-019-0195-x

[7] Schenck RJ, Miller FH, Keswani RN. The Surveillance patterns of incidentally detected pancreatic cysts vary widely and infrequently adhere to guidelines. Pancreas 2019; 48: 883–887 doi:10.1097/MPA.0000000000001352

[8] Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Physical Syst 2023; 3: 121–154

[9] Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023; 6: 75 doi:10.3389/frai.2023.1169595

[10] Gorelik Y, Ghersin I, Maza I et al. Harnessing language models for streamlined postcolonoscopy patient management: a novel approach. Gastrointest Endosc 2023; 98: 639–641.e4

[11] Benoit JRA. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. medRxiv 2023; 2023.02.04.23285478. doi:10.1101/2023.02.04.23285478

[12] Okasha HH, Awad A, El-Meligui A et al. Cystic pancreatic lesions, the endless dilemma. World J Gastroenterol 2021; 27: 2664–2680 doi:10.3748/wjg.v27.i21.2664

[13] Marchegiani G, Pollini T, Burelli A et al. Surveillance for presumed BD-IPMN of the pancreas: Stability, size, and age identify targets for discontinuation. Gastroenterology 2023; 165: 1016–1024 doi:10.1053/j.gastro.2023.06.022