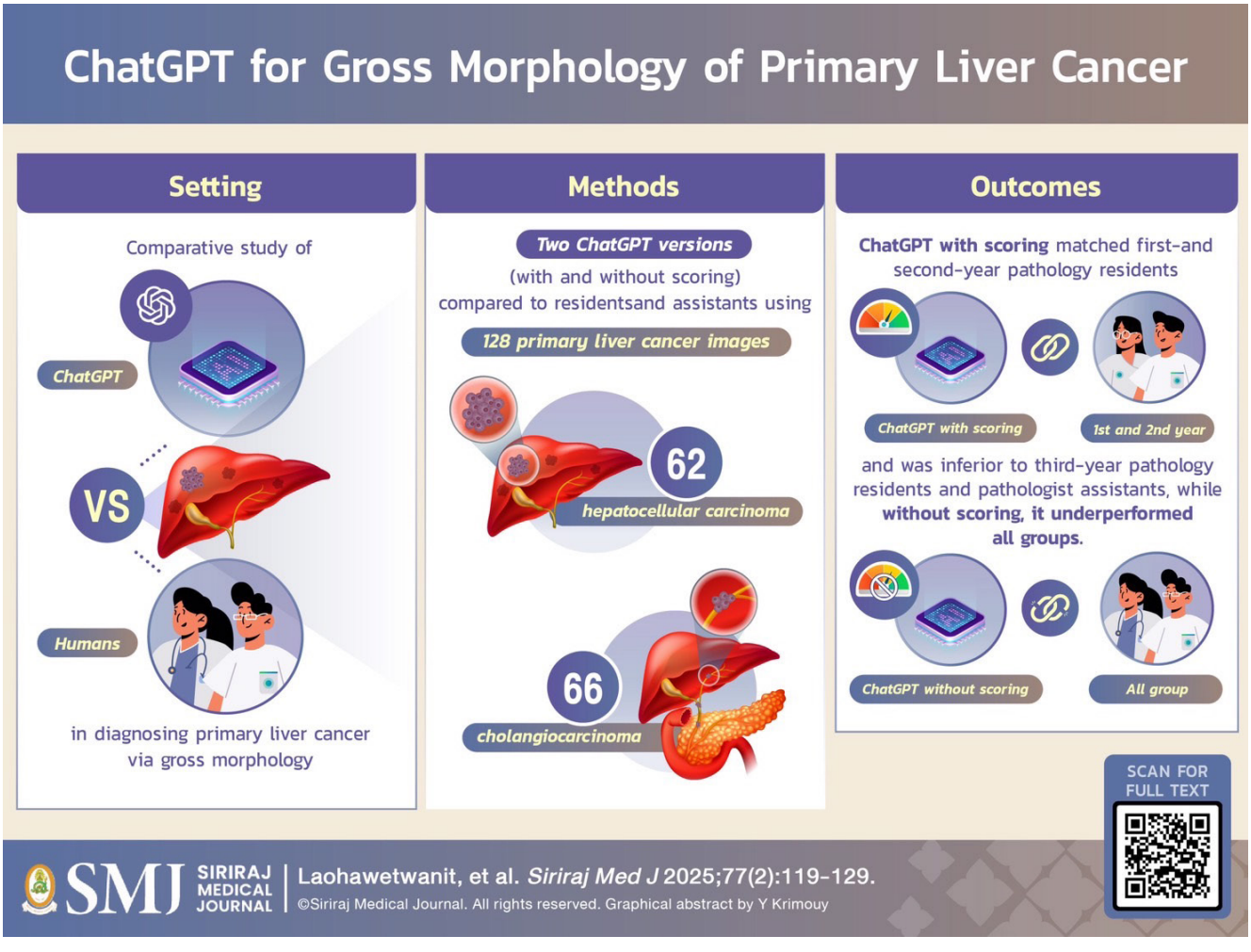
Comparative Analysis of ChatGPT and Human Expertise in Diagnosing Primary Liver Carcinoma: A Focus on Gross Morphology

Prakasit Sa-ngiamwibool, M.D.[1], Thiyaphat Laohawetwanit, M.D.[2,*]

[1]Department of Pathology, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand, [2]Division of Pathology, Chulabhorn International College of Medicine, Thammasat University, Pathum Thani, Thailand.



*Corresponding author: Thiyaphat Laohawetwanit
E-mail: thiyapat@tu.ac.th
Received 11 October 2024 Revised 15 November 2024 Accepted 15 November 2024
ORCID ID:http://orcid.org/0000-0003-3805-7291
https://doi.org/10.33192/smj.v77i2.271596

All material is licensed under terms of the Creative Commons Attribution 4.0 International (CC-BY-NC-ND 4.0) license unless otherwise stated.

ABSTRACT

Objective: This study aims to compare the diagnostic accuracy of customized ChatGPT and human experts in identifying primary liver carcinoma using gross morphology.

Materials and Methods: Gross morphology images of hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA) cases were assessed. These images were analyzed by two versions of customized ChatGPT (e.g., with and without a scoring system), pathology residents, and pathologist assistants. The diagnostic accuracy and consistency of each participant group were evaluated.

Results: The study analyzed 128 liver carcinoma images (62 HCC, 66 CCA), with the participation of 13 pathology residents (median experience of 1.5 years) and three pathologist assistants (median experience of 5 years). When augmented with a scoring system, ChatGPT's performance was found to align closely with first- and second-year pathology residents and was inferior to third-year pathology residents and pathologist assistants, with statistical significance (p-values < 0.01). In contrast, the diagnostic accuracy of ChatGPT, when operating without the scoring system, was significantly lower than that of all human participants (p-values < 0.01). Kappa statistics indicated that the diagnostic consistency was slight to fair for both customized versions of ChatGPT and the pathology residents. It was noted that the interobserver agreement among the pathologist assistants was moderate.

Conclusion: The study highlights the potential of ChatGPT for augmenting diagnostic processes in pathology. However, it also emphasizes the current limitations of this AI tool compared to human expertise, particularly among experienced participants. This suggests the importance of integrating AI with human judgment in diagnostic pathology.

Keywords: Artificial intelligence; ChatGPT; GPT-4; Liver cancer; Hepatocellular carcinoma; Cholangiocarcinoma
(Siriraj Med J 2025; 77: 119-129)

INTRODUCTION

Primary liver carcinoma, encompassing hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA), represents a significant global health challenge due to its high incidence and mortality rates.[1] Pathologists play a crucial role in diagnosing these malignancies, relying heavily on a combination of histopathological examination and the assessment of gross morphological features. This assessment is vital, as the initial diagnosis often depends on the visual and tactile examination of liver tissue during surgical procedures or post-surgical analysis. The gross morphology of liver carcinoma can provide critical insights into the tumor's type, size, extent, and potential invasiveness, guiding further histological examination. Accurate identification and characterization of these cancers at the gross level are essential, as they directly influence treatment decisions, prognostic evaluations, and patient outcomes.[2,3] Therefore, the precision and expertise brought by pathologists in interpreting these gross morphological features are of paramount importance in the clinical management of primary liver carcinoma.

The uses of artificial intelligence (AI) in medicine have been growing. A chatbot, such as ChatGPT, developed by OpenAI, combines a general-purpose AI system with a chat interface, allowing for natural conversations akin to those between humans. Users interact with ChatGPT by entering prompts, to which it responds contextually and quickly. The effectiveness of these interactions often depends on the precise wording of prompts, a concept known as "prompt engineering". ChatGPT excels in generating accurate responses to straightforward questions.[4] Evidence highlights ChatGPT-4's high success rate in answering USMLE practice questions and its superior performance compared to earlier AI chatbot models.[5] However, this generative AI program demonstrates promising yet limited capabilities in pathology education when being asked to answer board-style questions.[6]

The recent integration of image analysis capabilities into ChatGPT marks a significant advancement in the field of generative AI. This feature enables ChatGPT to comprehend and respond to textual inputs and analyze and interpret images. Users can now upload images directly into the conversation, allowing the AI to provide detailed descriptions, insights, and interpretations based on the visual content.[7] Data on using ChatGPT for image analysis in anatomical pathology is limited. This study aimed to compare the accuracy of the diagnosis of primary liver carcinoma based on gross morphology between customized versions of ChatGPT and human expertise.

MATERIALS AND METHODS

Gross images

Gross images of consecutive cases of HCC and CCA were obtained from the archives of the Department of Pathology, Faculty of Medicine, Khon Kaen University, capturing a broad spectrum of morphological variations in HCC and CCA. Diagnoses were confirmed by a liver pathologist (PS) through histopathologic evaluation using hematoxylin and eosin (H&E)-stained slides, with immunohistochemistry as an adjunct, to establish ground truth. Clinical data for each case was collected from an electronic medical database.

ChatGPT-4

Scoring system

ChatGPT-4 was assigned the task of developing a scoring system to differentiate the gross morphological features of HCC and CCA using the prompt, "Create a scoring system to distinguish gross morphology between HCC and CCA". The initial system included seven parameters: lesion color, texture, nodule presentation, vascular invasion, capsule formation, growth pattern, and presence of satellite lesions. Each parameter was scored on a scale from 0 to 2, representing typical features of HCC and CCA. After a review by two hepatobiliary pathologists, PS and TL, the scoring system was refined, with vascular invasion removed, leaving six criteria. The updated system was validated using 20 images of HCC and CCA.

Customization

For a uniform approach to evaluating image responses, we used two distinct configurations of ChatGPT-4. The first configuration applied a predefined scoring system before arriving at a final diagnosis. In contrast, the second configuration directly determined the final diagnosis, bypassing the scoring system. Snapshots of these configurations are presented in Fig 1.

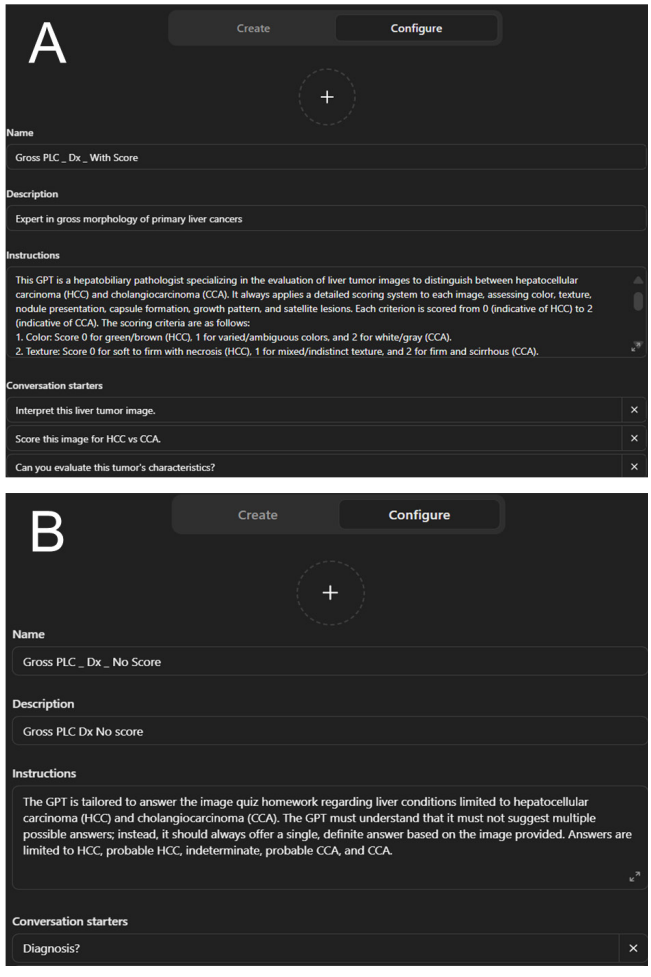


Fig 1. Snapshots of ChatGPT's configurations. (A) With scoring system. (B) Without scoring system.

Detailed guidelines for both configurations of ChatGPT-4 are provided in the following sections.

1. ChatGPT with scoring system

"This GPT is a hepatobiliary pathologist specializing in the evaluation of liver tumor images to distinguish between hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA). It always applies a detailed scoring system to each image, assessing color, texture, nodule presentation, capsule formation, growth pattern, and satellite lesions. Each criterion is scored from 0 (indicative of HCC) to 2 (indicative of CCA). The scoring criteria are as follows:

1. Color: Score 0 for green/brown (HCC), 1 for varied/ambiguous colors, and 2 for white/gray (CCA).

2. Texture: Score 0 for soft to firm with necrosis (HCC), 1 for mixed/indistinct texture, and 2 for firm and scirrhous (CCA).

3. Nodule presentation: Score 0 for single/multiple nodules (HCC), 1 for multiple nodules with unclear margins, and 2 for absence of nodules (CCA).

4. Capsule formation: Score 0 for the presence of a fibrous capsule (HCC), 1 for partial/incomplete capsule, and 2 for no capsule (CCA).

5. Growth pattern: Score 0 for expansive growth (HCC), 1 for mixed/uncertain growth patterns, and 2 for infiltrative growth (CCA).

6. Satellite lesions: Score 0 for numerous satellite lesions (HCC), 1 for a moderate number, and 2 for absent/few (CCA).

The score interpretation is as follows: 0-2 indicates definite HCC, 3-5 suggests probable HCC, 6 is indeterminate, 7-9 points towards probable CCA, and 10-12 indicates definite CCA. This method ensures a thorough and accurate analysis for each case."

2. ChatGPT without scoring system

"The GPT is tailored to answer the image quiz homework regarding liver conditions limited to hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA). The GPT must understand that it must not suggest multiple possible answers; instead, it should always offer a single, definite answer based on the image provided. Answers are limited to HCC, probable HCC, indeterminate, probable CCA, and CCA".

To evaluate the diagnostic consistency of ChatGPT, each image was sent to ChatGPT for 5 times. The responses from ChatGPT for each submission of the image were then recorded.

Human participants

Pathology residents and pathologist assistants at the Department of Pathology, Faculty of Medicine, Khon Kaen University, were tasked with diagnosing primary liver carcinoma based on gross images. While not briefed on ChatGPT-4's specific scoring system, these participants were instructed to classify each image according to categories derived from ChatGPT-4's analysis. These categories included definite HCC, probable HCC, indeterminate, probable CCA, and definite CCA.

Marking system

The accuracy of ChatGPT's responses and those of human participants were assessed against ground truths (e.g., histopathological evaluation). A marking system was established to measure the proficiency of ChatGPT and human experts in diagnosing primary liver carcinoma (Table 1).

TABLE 1. Marking system to measure the proficiency of ChatGPT and human experts in diagnosing primary liver carcinoma.

Ground truth	Response	Mark
HCC	Definite HCC	2
	Probable HCC	1
	Indeterminate	0
	Probable CCA	-1
	Definite CCA	-2
CCA	Definite CCA	2
	Probable CCA	1
	Indeterminate	0
	Probable HCC	-1
	Definite HCC	-2

Abbreviations: HCC, hepatocellular carcinoma; CCA, cholangiocarcinoma.

Sample size justification

A sample size calculation was conducted to determine the number of images required per group for reliable comparisons of diagnostic accuracy between ChatGPT and human evaluators. Using an independent t-test with a significance level of 0.05, a power of 80%, and a moderate effect size (0.5), the analysis indicates that approximately 64 images per group are needed. This sample size provides an 80% probability of detecting a true difference in diagnostic accuracy if one exists, with a 5% chance of a Type I error (false positive).

Data analysis

Descriptive statistics were used to detail the gross image and characteristics of the participants. The performance of ChatGPT-4 and human experts was assessed through overall scores. The scores were compiled for each entity: ChatGPT with a scoring system, ChatGPT without a scoring system, residents from each year of their program, and pathologist assistants. The ANOVA test was used to calculate whether there were differences among the groups. To provide insights into the specific groups between which significant differences existed, Tukey's Honestly Significant Difference (HSD) test was used for post-hoc analysis. A p-value of less than 0.05 indicated a statistically significant difference. Interobserver agreement was assessed with kappa statistics, interpreted according to the Landis and Koch scale: values <0.00 indicate poor agreement, 0.00–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement.[8]

RESULTS

Gross image and participant characteristics

The study included 128 macroscopic images of primary liver carcinoma, consisting of 62 cases of HCC and 66 cases of CCA. Thirteen pathology residents (5 first-year, 3 second-year, and 5 third-year residents) and three pathologist assistants participated in the analysis. The residents had a median of 1.5 years of pathology experience (range: 0.5 to 2.5 years) but lacked specific training in liver pathology and had limited exposure to primary liver carcinoma cases, with no recent educational activities that might have influenced their diagnostic performance. In contrast, the pathologist assistants had a median of 5 years of experience (range: 3 to 7 years), were trained in handling liver specimens, and were familiar with the gross morphology of primary liver carcinomas.

Performance of ChatGPT-4 and human participants

Snapshots of ChatGPT's performance are demonstrated in Fig 2. Participants' scores are illustrated in Fig 3. The performance of each participant group is summarized in Table 2. The customized ChatGPT that did not apply a scoring system performed worst among all evaluators. The highest mean scores were those of pathologist assistants and third-year residents.

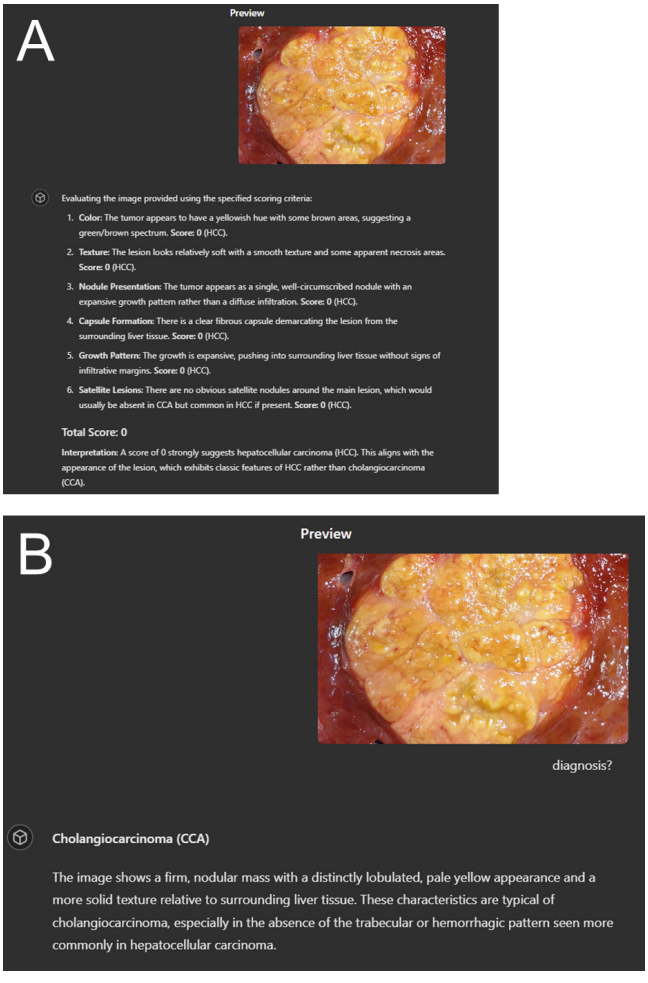


Fig 2. Snapshots of ChatGPT's responses. (A) With scoring system. (B) Without scoring system.

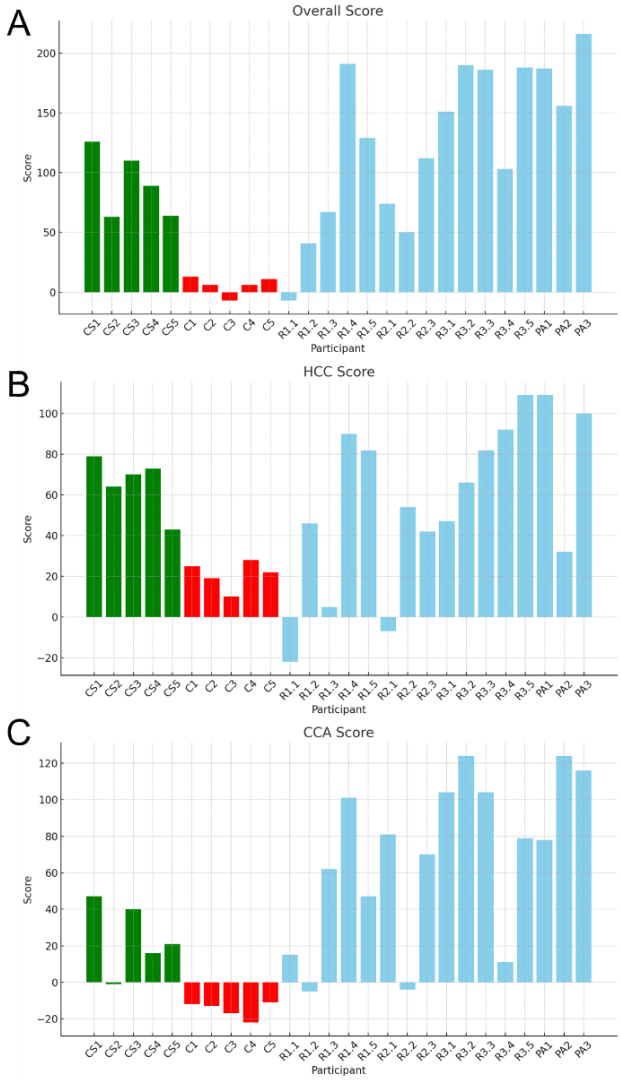


Fig 3. Scores of each participant. Bar charts demonstrating (A) overall scores, (B) hepatocellular carcinoma (HCC) scores, and (C) cholangiocarcinoma (CCA) score.
Abbreviations: CS = ChatGPT with a scoring system; C = ChatGPT without a scoring system; R1 = First-year pathology residents; R2 = Second-year pathology residents; R3 = Third-year pathology residents; PA = Pathologist assistants.

TABLE 2. Performance of ChatGPT 4 and human participants.

Evaluators	Mean	SD	Min	Max
ChatGPT				
Scoring system	90.4	24.9	63	123
No scoring system	5.8	7	-7	13
Human participants				
First-year residents	84.2	69.1	-7	191
Second-year residents	78.7	25.5	50	112
Third-year residents	163.6	33.5	103	190
Pathologist assistants	186.3	24.5	156	216

Abbreviations: SD = Standard deviation; Min = Minimum; Max = Maximum.

Comparison of scores among ChatGPT-4 and human participants

The Shapiro-Wilk test and Levene's test were applied to assess normality and homogeneity of variances, respectively, yielding p-values above 0.05 before performing the ANOVA test. The ANOVA showed an F-statistic of 9.804 and a p-value below 0.01. A post-hoc Tukey's HSD test was then conducted to identify specific group differences (Table 3). In scoring comparisons, ChatGPT's average score was similar to those of first- and second-year pathology residents but was lower than those of third-year residents and pathologist assistants (p-value < 0.01). Notably, when ChatGPT operated without a scoring system, its overall performance was significantly lower than that of all human participant groups (p-value < 0.01). The data also indicated a trend of higher scores among human participants as their pathology experience increased.

A subgroup analysis focusing on comparing the diagnostic abilities of HCC and CCA among different participant groups was performed. In assessing the HCC diagnostic scores, the ANOVA test revealed an F-statistic of 3.10 and a p-value of 0.03. Despite this, none of the group-to-group comparisons showed statistically significant differences in average scores, as all p-values exceeded 0.05. This suggests that while there was some overall variance among the groups, no single group differed significantly.

Regarding CCA diagnostic scores, the ANOVA test indicated a more pronounced difference with an F-statistic of 7.22 and a p-value below 0.01. Detailed pairwise comparisons revealed significant score differences between several groups: ChatGPT with a scoring system and pathologist assistants (p-value = 0.03), ChatGPT without a scoring system and pathologist assistants (p-value < 0.01), and ChatGPT without a scoring system and third-year pathology residents (p-value < 0.01). Conversely, the remaining group comparisons did not show statistically significant differences.

TABLE 3. Multiple comparison of means by the Tukey's HSD test.

Group 1	Group 2	Mean difference	P-value	Lower CI	Upper CI	Reject
CS	C	-84.60	<0.01*	-115.59	-53.61	TRUE
CS	PA	95.93	<0.01*	64.95	126.92	TRUE
CS	R1	-6.20	0.9	-37.19	24.79	FALSE
CS	R2	-11.73	0.9	-42.72	19.26	FALSE
CS	R3	73.20	<0.01*	42.21	104.19	TRUE
C	PA	180.53	<0.01*	149.55	211.52	TRUE
C	R1	78.40	<0.01*	47.41	109.39	TRUE
C	R2	72.87	<0.01*	41.88	103.86	TRUE
C	R3	157.80	<0.01*	126.81	188.79	TRUE
PA	R1	-102.13	<0.01*	-133.12	-71.14	TRUE
PA	R2	-107.67	<0.01*	-138.66	-76.68	TRUE
PA	R3	-22.73	0.69	-53.72	8.26	FALSE
R1	R2	-5.53	0.9	-36.52	25.46	FALSE
R1	R3	79.40	<0.01*	48.41	110.39	TRUE
R2	R3	84.93	<0.01*	53.95	115.92	TRUE

The table shows the mean differences between the scores of each pair of groups, along with the p-value for each comparison. The lower and upper bounds of the 95% confidence interval for the mean difference are presented. If the "reject" column is true, it indicates that there is a statistically significant difference between those groups.

Abbreviations: CS = ChatGPT with scoring system; C = ChatGPT without scoring system; PA = Pathologists assistants; R1 = First-year pathology residents; R2 = Second-year pathology residents; R3 = Third-year pathology residents; CI = Confidence interval.

* = Statistically significant difference.

Diagnostic consistency

Heat maps showing diagnostic consistency among participants are presented in Fig 4. Kappa statistics with interpretations of agreement level are shown in Table 4. Similar to pathology residents, both customized ChatGPT versions provided slight to fair diagnostic consistency. Notably, the level of agreement in the pathologist assistant group was moderate, higher than that of other participant groups. A subgroup analysis was conducted to compare the diagnostic consistency for HCC and CCA across different participant groups. While the use of ChatGPT without a scoring system resulted in moderate diagnostic consistency, other groups demonstrated only slight agreement in diagnosing both HCC and CCA.

Qualitative analysis of images with consistent responses

Images that consistently received definitive diagnoses from participants were subjected to further analysis. Out of these, 35 images (56.5%) of HCC and 20 images (30.3%) of CCA were classified as "definite" for HCC and CCA, respectively, by over half of the participants. Specifically, the three HCC images that received the highest number of "definite" diagnoses (ranging from 18 to 20 out of 26 responses) depicted either a mass in a cirrhotic background (Fig 5A) or a solitary nodule exhibiting expansive growth (Fig 5B). Similarly, the three CCA images with the most definitive diagnoses (16 to 21 responses) displayed characteristics such as an intraductal growth pattern (Fig 5C) or a combination of periductal infiltrating and mass forming growth patterns (Fig 5D). The HCC image that most frequently received reversed diagnoses of CCA (12 responses) presented a large, protruding mass with a minimal amount of non-neoplastic liver tissue (Fig 5E). Conversely, the CCA image most commonly mistaken for HCC (11 responses) predominantly showed a mass forming growth pattern with focal areas of periductal infiltration (Fig 5F).

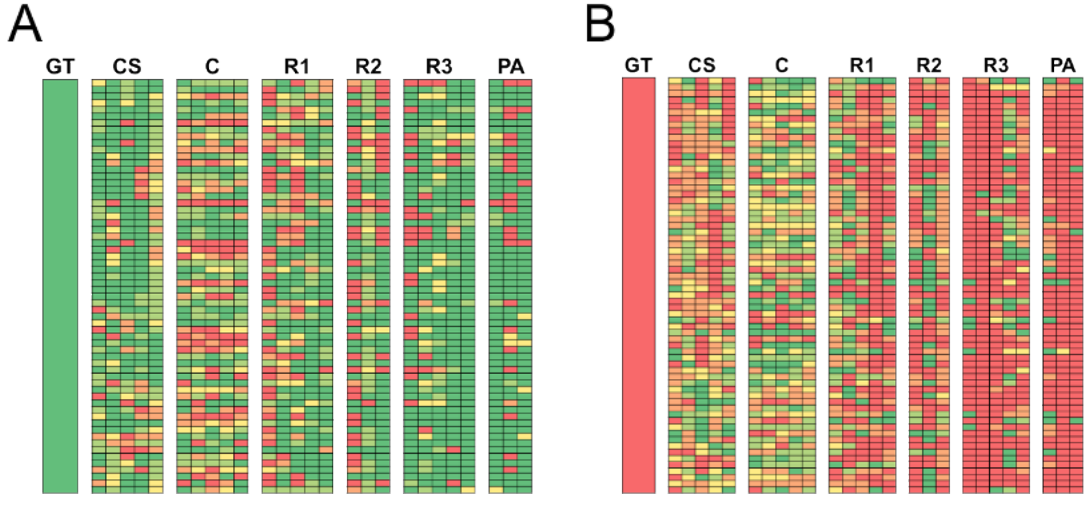


Fig 4. Diagnostic consistency. Heat maps showing diagnostic consistency of (A) hepatocellular carcinoma (HCC) and (B) cholangiocarcinoma (CCA) among each participant group. Different colors were used to show diagnostic confidence levels for HCC and CCA as follows: green for definite HCC, light green for probable HCC, yellow for indeterminate, orange for probable CCA, and red for definite CCA.
Abbreviations: GT = Ground truth; CS = ChatGPT with a scoring system; C = ChatGPT without a scoring system; R1 = First-year pathology residents; R2 = Second-year pathology residents; R3 = Third-year pathology residents; PA = Pathologist assistants.

TABLE 4. Diagnostic consistency of each participant group.

Participant group	Kappa	Level of agreement
ChatGPT with a scoring system	0.07	Slight
ChatGPT without a scoring system	0.25	Fair
First-year residents	0.06	Slight
Second-year residents	0.02	Slight
Third-year residents	0.31	Fair
Pathologist assistants	0.46	Moderate

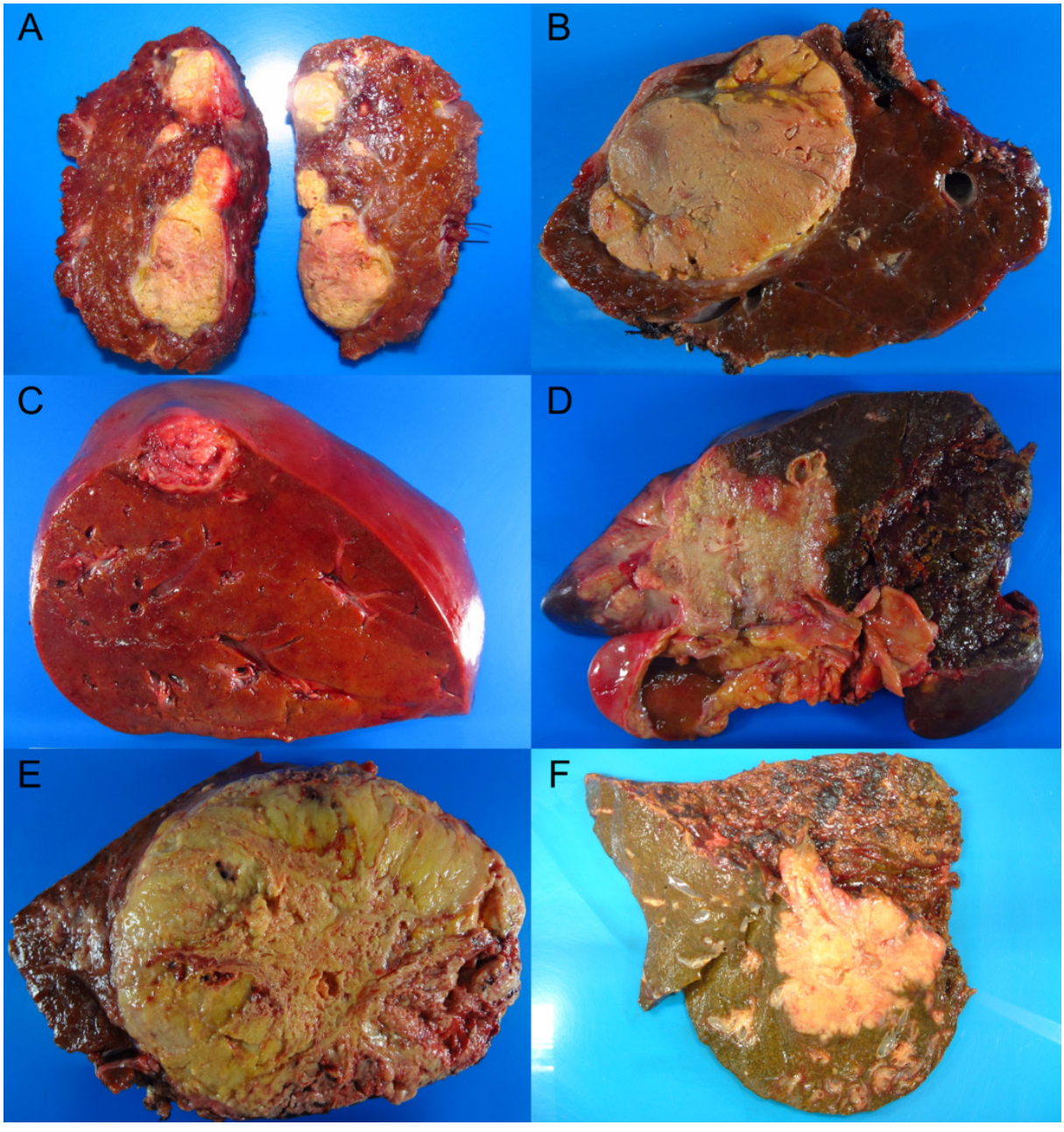


Fig 5. Images with consistent responses. Well-demarcated masses in a cirrhotic background (A) and a single nodule with an expansive growth (B) are usually diagnosed as hepatocellular carcinoma. Masses with intraductal (C) and unequivocal periductal infiltrating (D) growth patterns are commonly regarded as cholangiocarcinoma. A large, bulging mass with limited non-neoplastic liver tissue is sometimes mistaken for cholangiocarcinoma (E). Cholangiocarcinoma with a predominant mass-forming growth can be misidentified as hepatocellular carcinoma (F).

DISCUSSION

Our study provides a comparative analysis of ChatGPT and human experts in diagnosing primary liver carcinoma based on gross morphology. Key findings reveal that both versions of ChatGPT—one using a scoring system and one without—showed limited accuracy and consistency in diagnosing primary liver carcinoma through gross morphology alone. ChatGPT with a scoring system performed at a level similar to that of first- and second-year pathology residents but was less effective than more experienced residents and pathologist assistants. Notably, ChatGPT without a scoring system had the lowest performance among all evaluators. The slight to fair agreement observed among ChatGPT versions is inadequate for practical application, contrasting with the moderate agreement level among pathologist assistants with experience in gross morphology of primary liver carcinoma. These findings highlight the current limitations of AI in interpreting gross morphology in neoplastic diseases, underscoring the essential role of human expertise in this domain.

As of this writing, there are limited data on ChatGPT's use for analyzing pathological images, and the available literature shows inconsistent results. While ChatGPT-4 demonstrates high sensitivity in identifying adenomas, it struggles with low specificity, and its performance in polyp classification remains variable. One major concern is its lack of diagnostic consistency.[9] Additionally, its histopathological evaluations for common diseases are less accurate than those of pathology trainees.[10] However, applying step-by-step reasoning (also known as chain-of-thought prompting) has been shown to improve its performance.[11] In one study, ChatGPT was able to accurately grade and stage fatty liver disease using H&E and Masson's trichrome-stained images. Yet, caution is needed when interpreting these results, as they relied on a small sample of publicly available images (e.g., from PathologyOutlines.com) during a time when ChatGPT had internet browsing capabilities.[12,13] Another study reports that ChatGPT outperformed expert technicians in distinguishing between normal and abnormal blood cells.[14] However, the applicability of these findings is constrained by the specific types of images (e.g., histopathology, peripheral blood smears) utilized in the research. Despite high expectations among pathologists for AI chatbots to excel in image analysis,[15] these expectations has yet to be realized. The current study suggests that integrating a scoring system could improve ChatGPT-4's diagnostic accuracy.

ChatGPT, a cutting-edge chatbot developed by OpenAI, stands at the forefront of natural language processing technology.[16] Built on the sophisticated GPT-4 architecture, this chatbot is adept at generating text that closely resembles human conversation, making it incredibly versatile and user-friendly. Its training involves extensive analysis of diverse text data, enabling it to perform a variety of tasks such as providing informed responses, drafting emails, generating creative content, and even programming.[17] In 2023, ChatGPT underwent significant advancements, most notably the introduction of image processing. These features, enabled by GPT-4V and updated DALL-E models, allowed users to engage with ChatGPT through images, expanding its applications beyond text.[7] Additionally, the custom instructions feature was introduced, allowing users to personalize their interactions with ChatGPT to meet specific needs.[18] These developments not only enhanced user experience but also increased the tool's adaptability across various industries, highlighting its growing importance in AI-driven solutions and its potential for widespread impact in different sectors.

Research has demonstrated that ChatGPT-4 effectively answers a significant proportion of practice questions for the USMLE examinations[5,19] and offers accurate and consistent responses to materials designed for pathology examination preparation.[20] Additionally, a study has shown that ChatGPT can respond to pathology-specific inquiries on par with or surpass the performance of a trained pathologist, depending on the version of the model (GPT-3.5 or GPT-4).[21] Despite these findings, it is crucial to recognize that these outcomes do not imply ChatGPT's equivalence to pathologists in diagnostic abilities. The studies specifically exclude questions that require the interpretation of images. Furthermore, the nature of the examination questions, which tend to have clear and definitive answers, differs significantly from real-world diagnostic scenarios that demand complex decision-making.[22]

In a comparative study between ChatGPT and pathologists regarding histopathology diagnosis and collaborative potential, ChatGPT's contributions were deemed somewhat helpful but their effectiveness varied, contingent on the pathologists' prompts.[23] This finding highlights the importance of prompt engineering. Prompt engineering is essential in the AI field, particularly for advanced language models ChatGPT. It involves crafting questions or instructions to elicit precise and relevant responses from AI. Effective prompts also help mitigate misinterpretation and bias, leading to more reliable AI outcomes.[24]

Our study introduces a novel method for assessing primary liver carcinoma by comparing ChatGPT's image analysis capabilities with human pathologist expertise. A key strength lies in the innovative use of generative AI to create a scoring system, subsequently reviewed by liver pathologists—a first in this field. Similar research in radiology, applying AI to evaluate gross tumor volume and grade prognostic assessments of HCC, has shown promising results.[25,26] Additionally, the study pioneers AI analysis of gross morphological images of liver carcinomas, directly comparing AI and human assessments. This approach highlights AI's potential to support diagnostic processes in pathology while offering valuable insights into AI's current capabilities and limitations in medical image interpretation. By including both AI and human evaluators with diverse experience levels and no standardized training, the research methodology provides a realistic comparison, emphasizing areas where AI can complement human judgment and where human expertise remains indispensable.

While this study introduces a novel approach, several limitations affect its broader applicability. First, its focus on the gross morphology of primary liver carcinomas

may restrict the findings' relevance to other diseases and histopathologic evaluations. Incorporating microscopic findings, expert-provided descriptions, or other AI models might yield more promising outcomes.[10] Second, as the study was conducted within a single institution, the specialized expertise of pathology residents and pathologist assistants in cholangiocarcinoma may have influenced the results. Third, ChatGPT's current lack of annotation capabilities in image analysis could have impacted diagnostic precision; with the tool evolving rapidly, these findings may soon become outdated. Additionally, using other commercially available generative AI models, such as Gemini (formerly Bard) (Google, California, USA) and Claude (Anthropic, California, USA), might yield different outcomes. However, a recently published study has suggested that Bard underperformed compared to ChatGPT in pathology image interpretation.[27] These factors limit the generalizability of the study and highlight areas for further research and improvement.

This study offers several implications for pathology practice and future research. First, it highlights ChatGPT's potential to enhance human expertise in pathology, especially in analyzing and interpreting gross morphological features. ChatGPT's development of a scoring system represents a step toward integrating AI into diagnostic frameworks. Future research could benefit from advancing AI-based diagnostic criteria that incorporate both gross and microscopic features for more comprehensive diagnoses. Another promising application is using ChatGPT to suggest or interpret immunohistochemistry in challenging cases.[28] Additionally, expanding research to encompass various neoplastic and non-neoplastic diseases and testing across different institutions with diverse disease prevalences could provide valuable insights into the broader applicability and effectiveness of AI tools in pathology. It is important not to overstate ChatGPT's potential. As a fundamentally text-based large language model, ChatGPT generally excels with text-based tasks, such as generating pathological diagnoses from detailed microscopic descriptions or creating educational content,[10,29] rather than with tasks requiring pure image interpretation.[9-11]

In conclusion, this study underscores the evolving role of AI in pathology, particularly in diagnosing primary liver carcinoma using gross morphology. While ChatGPT exhibits potential in augmenting diagnostic processes, its performance, particularly with a scoring system, remains inferior to experienced pathology residents and pathologist assistants. The findings emphasize the importance of combining AI tools with human judgment in diagnostic pathology.

Data Availability Statement

Data are available on reasonable request.

ACKNOWLEDGEMENTS

The authors would like to thank pathology residents and pathologist assistants at the Department of Pathology, Faculty of Medicine, Khon Kaen University, for providing diagnoses of images used in this study. The authors thank Mr. Suwit Balthaisong, Department of Pathology, Khon Kaen University for his skillful assistance gross images.

DECLARATION

Grants and Funding Information

This study received no funding support.

Conflict of Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Both authors involved in conception, design of the study, acquisition and analysis of data, drafting and reviewing the manuscript.

Use of Artificial Intelligence

ChatGPT-4 was utilized to refine the English language in this manuscript. However, the accuracy and validity of the content are solely the responsibility of the authors.

Ethics approval statement

Ethical approval was provided by the Khon Kaen University Ethics Committee for Human Research based on the Declaration of Helsinki (Number HE671129).

REFERENCES

1. Rumgay H, Ferlay J, de Martel C, Georges D, Ibrahim AS, Zheng R, et al. Global, regional and national burden of primary liver cancer by subtype. Eur J Cancer. 2022;161:108-18.
2. Fan Z, Jin M, Zhang L, Wang N, Li M, Wang C, et al. From clinical variables to multiomics analysis: a margin morphology-based gross classification system for hepatocellular carcinoma stratification. Gut. 2023;72(11):2149-63.
3. Zen Y. Intrahepatic cholangiocarcinoma: typical features, uncommon variants, and controversial related entities. Hum Pathol. 2023;132:197-207.
4. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023;388(13): 1233-9.
5. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach. 2023:1-7.

6. Geetha SD, Khan A, Khan A, Kannadath BS, Vitkovski T. Evaluation of ChatGPT pathology knowledge using board-style questions. Am J Clin Pathol. 2024;161(4):393-98.
7. OpenAI. GPT-4V(Ision) system card2023. Available from: https://cdn.openai.com/papers/GPTV_System_Card.pdf.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
9. Laohawetwanit T, Namboonlue C, Apornvirat S. Accuracy of GPT-4 in histopathological image detection and classification of colorectal adenomas. J Clin Pathol. 2024:jcp-2023-209304.
10. Apornvirat S, Thinpanja W, Damrongkiet K, Benjakul N, Laohawetwanit T. Comparing customized ChatGPT and pathology residents in histopathologic description and diagnosis of common diseases. Ann Diagn Pathol. 2024;73:152359.
11. Laohawetwanit T, Apornvirat S, Namboonlue C. Thinking like a pathologist: Morphologic approach to hepatobiliary tumors by ChatGPT. Am J Clin Pathol. 2024:aqae087.
12. Zhang Y, Liu H, Sheng B, Tham YC, Ji H. Preliminary Fatty Liver Disease Grading Using General-Purpose Online Large Language Models: ChatGPT-4 or Bard? J Hepatol. 2024;80(6): e279-81.
13. OpenAI. ChatGPT plugins2023. Available from: https://openai.com/blog/chatgpt-plugins.
14. Yang WH, Yang YJ, Chen TJ. ChatGPT's innovative application in blood morphology recognition. J Chin Med Assoc. 2024;87(4): 428-33.
15. Laohawetwanit T, Pinto DG, Bychkov A. A survey analysis of the adoption of large language models among pathologists. Am J Clin Pathol. 2024:aqae093.
16. OpenAI. Introducing ChatGPT2022. Available from: https://openai.com/blog/chatgpt.
17. OpenAI. GPT-4 Technical Report2023. Available from: https://arxiv.org/pdf/2303.08774.pdf.
18. OpenAI. Introducing GPTs2023. Available from: https://openai.com/blog/introducing-gpts.
19. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep. 2023;13(1):16492.
20. Koga S. Exploring the pitfalls of large language models: Inconsistency and inaccuracy in answering pathology board examination-style questions. Pathol Int. 2023;73(12):618-20.
21. Wang AY, Lin S, Tran C, Homer RJ, Wilsdon D, Walsh JC, et al. Assessment of Pathology Domain-Specific Knowledge of ChatGPT and Comparison to Human Performance. Arch Pathol Lab Med. 2024.
22. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq. 2023.
23. Oon ML, Syn NL, Tan CL, Tan KB, Ng SB. Bridging bytes and biopsies: A comparative analysis of ChatGPT and histopathologists in pathology diagnosis and collaborative potential. Histopathology. 2024;84(4):601-13.
24. Mesko B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 2023; 25:e50638.
25. Zhang HW, Huang DL, Wang YR, Zhong HS, Pang HW. CT radiomics based on different machine learning models for classifying gross tumor volume and normal liver tissue in hepatocellular carcinoma. Cancer Imaging. 2024;24(1):20.
26. Gross M, Haider SP, Ze'evi T, Huber S, Arora S, Kucukkaya AS, et al. Automated graded prognostic assessment for patients with hepatocellular carcinoma using machine learning. Eur Radiol. 2024;34(10):6940-52.
27. Apornvirat S, Namboonlue C, Laohawetwanit T. Comparative analysis of ChatGPT and Bard in answering pathology examination questions requiring image interpretation. Am J Clin Pathol. 2024;162(3):252-60.
28. Sukpanichnant S. Malignancy of the lymph node: How general practitioners and pathologists can achieve a definitive diagnosis. Siriraj Med J. 2022;74(9):604–17.
29. Laohawetwanit T, Apornvirat S, Kantasiripitak C. ChatGPT as a teaching tool: Preparing pathology residents for board examination with AI-generated digestive system pathology tests. Am J Clin Pathol. 2024;162(5):471-9.