Annals of Hepatology 30 (2025) 101537

Contents lists available at ScienceDirect

Annals of Hepatology

journal homepage: www.elsevier.es/annalsofhepatology

Original article

Evaluation of four chatbots in autoimmune liver disease: A comparative analysis

Jimmy Daza[a], Lucas Soares Bezerra[a], Laura Santamaría[b], Roberto Rueda-Esteban[b], Heike Bantel[c], Marcos Girala[d], Matthias Ebert[e,n], Florian Van Bömmel[f], Andreas Geier[g], Andres Gomez Aldana[h], Kevin Yau[i], Mario Alvares-da-Silva[j], Markus Peck-Radosavljevic[k], Ezequiel Ridruejo[l], Arndt Weinmann[m], Andreas Teufel[a,n,*]

[a] *Division of Hepatology, Division of Clinical Bioinformatics, Department of Medicine II, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*
[b] *Universidad de Los Andes School of Medicine, Bogotá, Colombia*
[c] *Department of Gastroenterology, Hepatology, Infectious Diseases and Endocrinology, Hannover Medical School, Hannover, Germany*
[d] *Department of Gastroenterology, Hospital de Clínicas, Universidad Nacional de Asunción, Asunción, Paraguay*
[e] *Department of Medicine II, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*
[f] *Department of Medicine II, Clinic of Gastroenterology, Hepatology, Infectious Diseases and Pneumology, Leipzig University Medical Center, Leipzig, Germany*
[g] *Department of Internal Medicine II, Division of Hepatology, University Hospital Würzburg, Würzburg, Germany*
[h] *Texas Liver Institute, University of Texas Health Science Center, San Antonio, United States*
[i] *Division of Nephrology, Department of Medicine, University of Toronto, Toronto, Ontario, Canada*
[j] *Department of Gastroenterology, Hospital de Clinicas de Porto Alegre, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil*
[k] *Internal Medicine and Gastroenterology (IMuG), Clinic Klagenfurt am Woerthersee, Klagenfurt, Austria*
[l] *Department of Medicine, Section of Hepatology, Centro de Educación Médica e Investigaciones Clínicas Norberto Quirno "CEMIC", Buenos Aires, Argentina*
[m] *Department of Internal Medicine I, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany*
[n] *Clinical Cooperation Unit Healthy Metabolism, Center for Digital Medicine and Prevention, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*

ARTICLE INFO

Article History:
Received 26 March 2024
Accepted 13 June 2024
Available online 13 August 2024

Keywords:
Artificial intelligence
Chatbots
Clinical decision support tools
Autoimmune liver disease

ABSTRACT

Introduction and Objectives: Autoimmune liver diseases (AILDs) are rare and require precise evaluation, which is often challenging for medical providers. Chatbots are innovative solutions to assist healthcare professionals in clinical management. In our study, ten liver specialists systematically evaluated four chatbots to determine their utility as clinical decision support tools in the field of AILDs.

Materials and Methods: We constructed a 56-question questionnaire focusing on AILD evaluation, diagnosis, and management of Autoimmune Hepatitis (AIH), Primary Biliary Cholangitis (PBC), and Primary Sclerosing Cholangitis (PSC). Four chatbots -ChatGPT 3.5, Claude, Microsoft Copilot, and Google Bard- were presented with the questions in their free tiers in December 2023. Responses underwent critical evaluation by ten liver specialists using a standardized 1 to 10 Likert scale. The analysis included mean scores, the number of highest-rated replies, and the identification of common shortcomings in chatbots performance.

Results: Among the assessed chatbots, specialists rated Claude highest with a mean score of 7.37 (*SD* = 1.91), followed by ChatGPT (7.17, *SD* = 1.89), Microsoft Copilot (6.63, *SD* = 2.10), and Google Bard (6.52, *SD* = 2.27). Claude also excelled with 27 best-rated replies, outperforming ChatGPT (20), while Microsoft Copilot and Google Bard lagged with only 6 and 9, respectively. Common deficiencies included listing details over specific advice, limited dosing options, inaccuracies for pregnant patients, insufficient recent data, over-reliance on CT and MRI imaging, and inadequate discussion regarding off-label use and fibrates in PBC treatment. Notably, internet access for Microsoft Copilot and Google Bard did not enhance precision compared to pre-trained models.

Abbreviations: AILD, autoimmune liver diseases; AIH, autoimmune hepatitis; PBC, primary biliary cholangitis; PSC, primary sclerosing cholangitis; AS-AIH, acute severe autoimmune hepatitis; LLM, large language models; GPT, Generative Pre-Trained Transformer; P-value, returned by chi-square test for independence; HRQL, health-related quality of life; MRI, magnetic resonance imaging; MRE, magnetic resonance enterography; CT, computed tomography; MRCP, magnetic resonance cholangiopancreatography; MELD, Mayo Clinic Model for End-Stage Liver Disease; MMF, mycophenolate mofetil; UDCA, ursodeoxycholic acid

* Corresponding author.
E-mail address: andreas.teufel@medma.uni-heidelberg.de (A. Teufel).

https://doi.org/10.1016/j.aohep.2024.101537

Conclusions: Chatbots hold promise in AILD support, but our study underscores key areas for improvement. Refinement is needed in providing specific advice, accuracy, and focused up-to-date information. Addressing these shortcomings is essential for enhancing the utility of chatbots in AILD management, guiding future development, and ensuring their effectiveness as clinical decision-support tools.

1. Introduction

Autoimmune liver diseases (AILDs) are rare, progressive conditions for which the precise triggers remain unknown, although genetic and environmental factors are considered potential contributors [1]. Autoimmune hepatitis (AIH), primary biliary cholangitis (PBC), and primary sclerosing cholangitis (PSC) are the most prevalent AILDs and vary in the target of immune-mediated injury, underlying pathophysiology, risk factors, and clinical phenotype. In AIH, hepatocytes are affected, while PBC and PSC involve the biliary system, specifically interlobular bile ducts, intra- and extrahepatic bile ducts, respectively [2].

The clinical presentation of AILDs is complex and consists of variable phenotypes, from low-risk, indolent courses to aggressive diseases, requiring precise evaluation and management. In approximately 10 % of cases, these diseases present as overlap variant syndromes in which both hepatocytes and cholangiocytes are injured by autoimmune response further complicating diagnosis and management [1].

Current studies suggest that AIH prevalence is rising globally, possibly related to lifestyle changes and improvements in diagnosis, with an estimated rate of 160–170 cases/100,000 worldwide and 200 cases/100,000 in North America [3]. The pooled global incidence and prevalence of PBC are estimated to range between 1.76 and 14.60 per 100,000 persons, respectively [4], while the incidence of PSC ranges from up to 1.3 per 100,000 inhabitants/year, and prevalence rates are up to 16.2 per 100,000 inhabitants [5]. Due to their rarity, AILDs may be unfamiliar to many medical providers. Delays in diagnosis or lack of knowledge surround appropriate treatment can lead to significant morbidity due to liver failure or development of hepatocellular carcinoma, consequently increasing the need for liver transplantation [6]. Notably, a review by Rahim *et al.* described that over 25 % of patients with acute severe autoimmune hepatitis (AS-AIH) present with acute liver failure [7].

The emergence of “Artificial Intelligence” chatbots based on Large Language Models (LLMs), offers the potential for innovative solutions to assist non-specialized healthcare professionals in improving the clinical management of rare diseases by leveraging the ability of chatbots to summarize large amounts of information in a precise and accurate manner, ideally improving physician efficiency [8]. A recent meta-analysis suggested that eHealth interventions may improve short-term medication adherence and self-monitoring behavior in solid organ transplant recipients [9]. Moreover, in a survey conducted regarding patients’ perspectives towards a deeper integration of eHealth in multi-disciplinary care, 38.7 % of the participants were convinced that further online communication in care could have a positive impact on physician-patient contact and 45.7 % assumed that eHealth use would have a positive impact on overall treatment quality [10].

Among the currently available LLMs, one of the most popular is ChatGPT, created by OpenAI. ChatGPT is modeled on Generative Pre-Trained Transformer (GPT), which is designed to process sequential data such as natural language, allowing it to better understand the context behind sentences, and to generate coherent text accordingly. Multiple publications suggest that ChatGPT may be used to improve the quality of medical education and clinical management, as it provides quick access to information, clinical decision support, and assists in communicating medical information with patients [11].

To address identify the limitations of AI and improving their utility in providing accurate medical information [12], our study systematically assessed the performance of four chatbots through a comprehensive questionnaire as evaluated by global liver specialists, and graded their utility as potential clinical decision support tools in the understanding, diagnosis and management of AILDs.

2. Materials and Methods

We conducted a comprehensive evaluation of four chatbots’ performance in addressing autoimmune liver diseases (AILD) – Autoimmune Hepatitis (AIH), Primary Biliary Cholangitis (PBC), and Primary Sclerosing Cholangitis (PSC). We developed a questionnaire comprising 50 questions focused on AILD’s General knowledge, up-to-date therapy, and two clinical decision cases. Following validation of the relevance of the questions in clinical practice, 6 additional questions were included (total of 56 questions). The questions were distributed by specific disease as follows: 20 for AIH, 18 for PBC, and 18 for PSC.

The chatbots included in the study were ChatGPT 3.5 by OpenAI (trained with data up to November 2022, https://chat.openai.com), Claude by Anthropic (trained with data up to December 2022, https://claude.ai), Microsoft Copilot by Microsoft (based on ChatGPT 4.0, with real-time Internet access, https://www.bing.com/chat), and Bard by Google (with real-time Internet access, https://bard.google.com/chat).

The queries presented to all chatbots were standardized: "I want you to act as an expert consultant doctor in hepatology and come up with precise replies for autoimmune liver diseases (AILD). I am a healthcare professional and will ask you certain specific concepts and management about AILD. Write every reply in under 300 words." In cases where the replies exceeded 300 words, an additional query was input: "Please, shorten the given text into a reply under 300 words."

The questions were introduced to each chatbot in their free and open-access subscription between December 10 and 20, 2023. Subsequently, responses were critically evaluated by a panel of 10 liver specialists, either head of their respective Hepatology departments or with over 5 years of experience as hepatology consultants, in Europe and the Americas. The panel consisted of 10 liver specialists, with 6 from Europe (5 in Germany, 1 in Austria) and 4 from the Americas (1 from Brazil, 1 from Argentina, 1 from Paraguay, and 1 from the USA). The evaluation, using a standardized 1 to 10 rating scale performed using a tailor-made online form using Fillout™ (www.fillout.com), occurred between January 10 and 20, 2024.

Comparative analysis was performed analyzing each chatbot’s mean score and identifying the number of highest-rated replies. Common shortcomings in chatbot performance in addressing complex hepatological queries were identified and described.

2.1. Ethical considerations

This study was a non-interventional study that analyzed the responses of chatbots to specific hepatology questions. As such, formal ethical approval from an Institutional Review Board or ethics committee was not required. This study adheres to ethical principles

of research, such as ensuring the protection of participant privacy and confidentiality, maintaining data security, and avoiding potential conflicts of interest or other ethical concerns.

3. Results

3.1. Comparative analysis

In a comprehensive examination, Claude demonstrated superior performance in analyzing the 56 questions regarding AILD, achieving a mean score of 7.37 (*SD* = 1.91), compared to 7.17 (*SD*= 1.89), 6.63 (*SD* = 2.10), and 6.52 (*SD* = 2.27) for ChatGPT, Copilot, and Bard, respectively. Additionally, when considering each specific disease, Claude outperformed other chatbots with mean scores of 7.38 for AIH, 7.27 for PBC, and 7.47 for PSC. The response quality according to each ChatBot is presented in Fig. 1.

In the AIH analysis, ChatGPT and Claude demonstrated highest scores for the three specific domains, with ChatGPT achieving the highest score on clinical decision questions, while Claude performed best for general questions and therapy (Table 1). Copilot had the lowest means in all general and clinical decision items. Bard had the two highest means when describing AIH and the histological changes found in the disease, however this was the only scenario in which Bard received a higher score than Copilot (6.98 vs 6.45, respectively).

For questions related to PBC (Table 2), scores were more equilibrated across chatbots in comparison with AIH. ChatGPT and Claude achieved similar means on general questions (7.39 and 7.45, respectively), although Claude had the highest scores in all three domains. Interestingly, for questions related to clinical decision making, Copilot had a higher mean in comparison with ChatGPT (6.70 vs 6.35).

For the analysis related to PSC, Claude performed best in all three domains, as shown in Table 3. With the exception of questions 8 and 16, where Bard and Copilot respectively achieved the highest scores, the highest rated replies for PSC were consistently provided by either ChatGPT or Claude. Notably, ChatGPT and Claude had identical mean scores in four questions: PSC-Q1 (7.90), PSC-Q6 (7.70), PSC-Q9 (7.30), and PSC-Q10 (7.30).

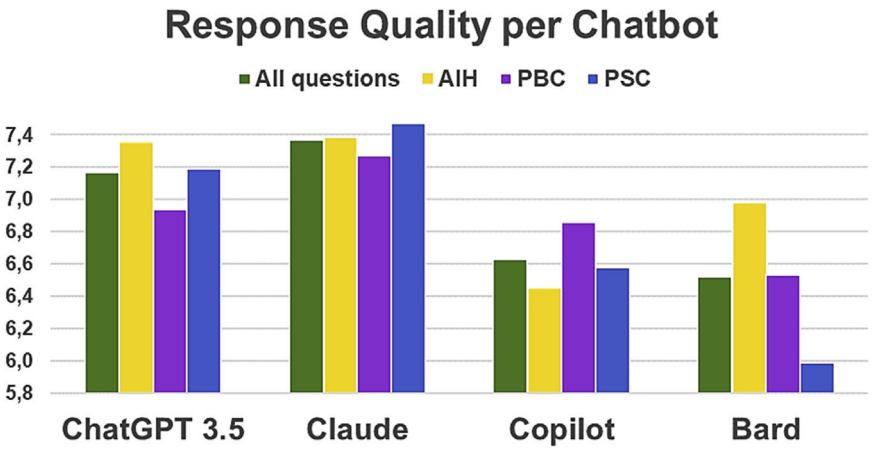

Fig. 1. Response quality per ChatBot in all questions, AIH, PBC, and PSC.

Table 1
Quality of answers provided by ChatGPT 3.5, Claude, Microsoft Copilot, and Google Bard related to autoimmune hepatitis.

	Chatbots			
Autoimmune Hepatitis (AIH)	ChatGPT	Claude	Copilot	Bard
General Questions	7.49	7.61	6.70	7.51
1. What is Autoimmune Hepatitis?	7.00	7.40	6.10	8.60
2. Which scores are used in the diagnosis of autoimmune hepatitis?	7.10	7.70	7.20	7.70
3. What histological changes are found in autoimmune hepatitis?	7.90	7.40	7.00	8.30
4. Are autoantibodies evidence of autoimmune hepatitis?	7.30	7.40	7.00	7.20
5. What other diseases can autoimmune hepatitis be associated with?	7.80	8.00	6.90	6.40
6. Can autoimmune hepatitis lead to liver cirrhosis?	7.90	7.70	6.50	7.70
7. What blood values do I need to take to diagnose autoimmune hepatitis?	7.70	7.40	6.70	7.20
8. What imaging do I need to diagnose autoimmune hepatitis?	7.20	7.90	6.20	7.00
Therapy Questions	7.11	7.15	6.38	6.54
9. What medications can be used to treat autoimmune hepatitis?	7.50	7.10	6.80	6.70
10. How can autoimmune hepatitis be treated during pregnancy?	7.60	7.30	7.50	4.70
11. What are the latest study results on autoimmune hepatitis?	5.00	6.00	6.40	6.80
12. What are the side effects of treating autoimmune hepatitis?	7.70	8.00	6.90	6.90
13. When should a patient with autoimmune hepatitis receive a liver transplant?	6.70	7.40	6.40	6.10
14. Should prednisolone, prednisone, or budesonide be used for first-line treatment of autoimmune hepatitis?	6.90	6.80	5.80	6.30
15. How quickly should cortisone be tapered off from therapy?	7.10	6.80	5.00	6.40
16. What is the preferred corticosteroid-sparing drug? Azathioprine?	7.30	7.70	6.30	6.60
17. If there is Azathioprine intolerance, what is the preferred second-line drug?	7.40	7.20	6.40	6.90
18. As a healthcare provider, when and how can I stop all treatment for autoimmune hepatitis?	7.90	7.20	6.30	8.00
Clinical Decision Questions	8.00	7.55	5.80	7.00
19. My patient with autoimmune hepatitis shows no improvement after cortisone administration. What should I do?	8.20	7.30	6.00	7.10
20. In my patient with autoimmune hepatitis, cortisone (or predniso(lo)ne) is not enough to normalize the liver values. What should I do?	7.80	7.80	5.60	6.90

Table 2
Quality of answers provided by ChatGPT 3.5, Claude, Microsoft Copilot, and Google Bard related to primary biliary cholangitis.

Primary Biliary Cholangitis (PBC)	Chatbots			
Autoimmune Hepatitis (AIH)	ChatGPT	Claude	Copilot	Bard
General Questions	7.39	7.45	7.14	7.18
1. What is Primary Biliary Cholangitis?	7.30	7.60	7.80	7.30
2. Which scores are used in the diagnosis of primary biliary cholangitis?	6.70	6.10	6.50	7.00
3. What histological changes are found in primary biliary cholangitis?	7.60	7.50	7.80	8.20
4. Are autoantibodies evidence of primary biliary cholangitis?	7.50	8.30	7.60	7.50
5. What other diseases can primary biliary cholangitis be associated with?	7.50	7.60	7.80	7.50
6. Can primary biliary cholangitis lead to liver cirrhosis?	7.70	7.90	6.80	6.80
7. What blood values do I need to take to diagnose primary biliary cholangitis?	8.00	7.70	7.20	7.30
8. What imaging do I need to diagnose primary biliary cholangitis?	6.80	6.90	5.60	5.80
Therapy Questions	6.65	7.13	6.63	6.15
9. What medications can be used to treat primary biliary cholangitis?	6.60	6.90	8.10	5.80
10. How can primary biliary cholangitis be treated during pregnancy?	7.00	6.50	6.00	6.10
11. What are the latest study results on primary biliary cholangitis?	6.10	6.50	6.60	5.30
12. What are the side effects of treating primary biliary cholangitis?	7.30	6.90	6.80	7.00
13. When should a patient with primary biliary cholangitis receive a liver transplant?	6.40	7.80	6.80	5.10
14. Should obeticholic acid or fibrates be used for second-line treatment of primary biliary cholangitis?	6.80	6.90	7.00	5.60
15. Should primary biliary cholangitis be treated with cortisone (predniso(lo)ne)?	7.30	7.90	5.10	6.40
16. Is it possible and recommended to stop treatment in primary biliary cholangitis?	5.70	7.60	6.60	7.90
Clinical Decision Questions	6.35	7.10	6.70	5.50
17. My patient with primary biliary cholangitis complains about severe itching. What should I do?	6.50	7.70	7.10	6.20
18. Can I give UDCA, obeticholic acid, and fibrate together in patients with Primary Biliary Cholangitis?	6.20	6.50	6.30	4.80

While Bard had the lowest mean scores in four out of five questions (AIH-Q10, PBC-Q18, PSC-Q13, and PSC-Q16) in the entire questionnaire, Copilot had the lowest mean score for any single question out of the 56 questions onPSC-Q14 (4.10), which involved describing the use of cortisone in PSC. Observations from evaluators suggested that Copilot did not adequately address the proposed question in PSC-Q14.

3.2. Evaluator's comments and considerations

Claude obtained the highest mean score in evaluations by seven out of 10 evaluators, while ChatGPT was preferred by the remaining three. Copilot and Bard were rated third and fourth by five evaluators each, as illustrated in Fig. 2. Additionally, Claude's responses were ranked first for 27 questions, whereas ChatGPT's responses were ranked highest for 20 questions. Bard and Copilot had the best-ranked replies for 9 and 6 questions, respectively, as illustrated in Fig. 3.

3.3. Evaluators' geographic influence

When considering the region of origin of the evaluators, significant differences were noticed between the European and American groups. Higher overall mean scores (8.02 vs. 6.19, as highlighted in the box of Fig. 2), as well as a smaller range of chatbot scores (0.44 vs 1.13), were reported in the Americas vs Europe respectively. When comparing individual chatbot scores, Claude was the highest rated in Europe, with a mean score of 6.81, while Claude and ChatGPT were tied for highest rate in the Americas, with a mean score of 8.21. Further results are summarized in Table 4.

Table 3
Quality of answers provided by ChatGPT 3.5, Claude, Microsoft Copilot, and Google Bard related to primary sclerosing cholangitis.

Primary Sclerosing Cholangitis (PSC)	Chatbots			
Autoimmune Hepatitis (AIH)	ChatGPT	Claude	Copilot	Bard
General Questions	7.44	7.64	6.74	6.49
1. What is Primary Sclerosing Cholangitis?	7.90	7.90	6.90	5.10
2. How is primary sclerosing cholangitis diagnosed?	7.00	7.90	6.90	6.40
3. What histological changes are found in primary sclerosing cholangitis?	7.90	7.20	7.00	7.40
4. Are autoantibodies evidence of primary sclerosing cholangitis?	7.10	7.90	5.30	6.50
5. What other diseases can primary sclerosing cholangitis be associated with?	7.40	7.30	7.20	6.30
6. Can primary sclerosing cholangitis lead to liver cirrhosis?	7.70	7.70	7.40	5.40
7. What blood values do I need to take to diagnose primary sclerosing cholangitis?	7.40	7.70	5.70	7.20
8. What imaging do I need to diagnose primary sclerosing cholangitis?	7.10	7.50	7.50	7.60
Therapy Questions	6.94	7.21	6.30	5.55
9. What medications can be used to treat primary sclerosing cholangitis?	7.30	7.30	6.40	5.10
10. How can primary sclerosing cholangitis be treated during pregnancy?	7.30	7.30	7.00	6.80
11. What are the latest study results on primary sclerosing cholangitis?	6.60	6.50	5.70	5.00
12. What are the side effects of treating primary biliary cholangitis?	6.70	5.40	5.20	6.40
13. When should a patient with primary biliary cholangitis receive a liver transplant?	6.40	8.00	7.50	4.70
14. Should primary biliary cholangitis be treated with cortisone (predniso(lo)ne)?	7.50	8.10	4.10	6.30
15. Should patients with primary sclerosing cholangitis receive scheduled ERCP?	6.70	7.50	6.40	5.50
16. Is screening for cholangiocarcinoma recommended? If so, how can I do it?	7.00	7.60	8.10	4.60
Clinical Decision Questions	7.25	7.85	7.10	5.75
17. My patient with primary sclerosing cholangitis keeps having fiber flare-ups. What should I do?	7.10	7.80	6.70	5.80
18. My patient with primary sclerosing cholangitis has liver cirrhosis. What should I do?	7.40	7.90	7.50	5.70

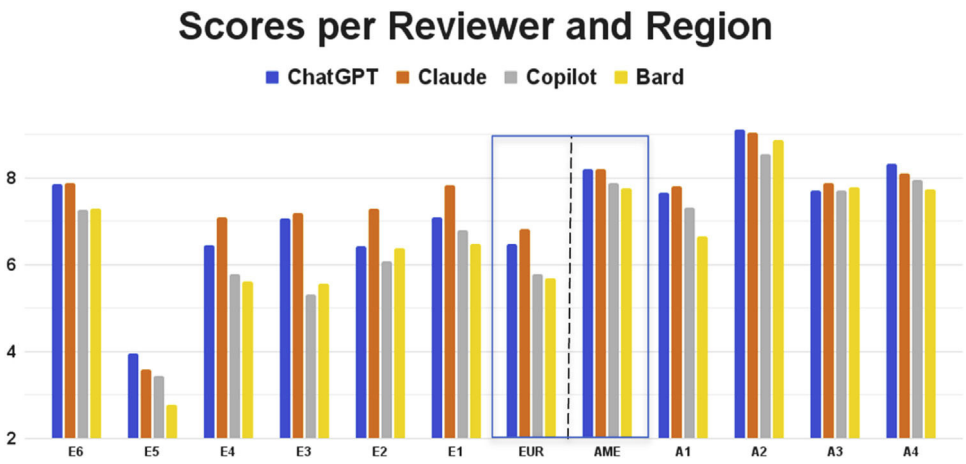

Fig. 2. Overall chatbot scores according to each evaluator. Please note that the box contains the mean score for the reviewers of the European (EUR) and Americas (AME) region.

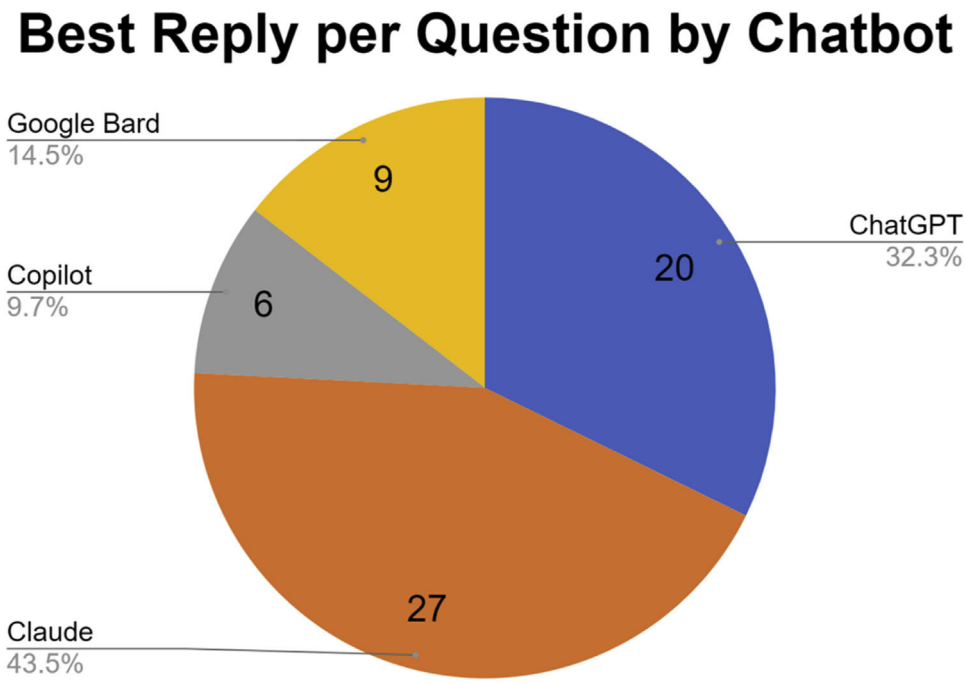

Fig. 3. Best reply per question by ChatBot, according to evaluators.

Table 4
Quality of answers provided by the chatbots overall and individually when considering the geographic origin of the evaluators.

	Europe	ChatGPT	Claude	Copilot	Bard	Americas	ChatGPT	Claude	Copilot	Bard
AIH	6.14	6.48	6.48	5.45	6.16	8.38	8.65	8.71	7.95	8.20
PBC	6.22	6.30	6.79	6.07	5.72	7.92	7.92	7.99	8.04	7.75
PSC	6.21	6.66	7.20	5.86	5.11	7.71	8.00	7.88	7.67	7.31
Overall	6.19	6.48	6.81	5.78	5.68	8.02	8.21	8.21	7.89	7.77

4. Discussion

Our study presents a comparison of the performance of four chatbots in assessing AILD. Our results indicate that Claude performed best overall and had an edge for both PBC and PSC, while ChatGPT performed the best for questions related to AIH.

To date, most of the published chatbot analyses related to liver disease have focused only on the performance of ChatGPT. This is a significant limitation considering that ChatGPT may lack information regarding regional variations in guidelines and, in some circumstances, may be inaccurate [13]. In our research, we aimed to improve this bias in research publications by including three other chatbots and specialists from multiple nationalities. In addition, this is the first study analyzing chatbot assessment on AILD.

Endo *et al.* found that ChatGPT's responses to liver transplant-related questions were highly accurate and complete, with over 70 %

graded as "Very Good" or "Excellent." This suggests ChatGPT's potential as a valuable resource for patient education and facilitating communication among medical professionals, patients, and caregivers [14]. Pugliese *et al.* evaluated ChatGPT's accuracy in addressing NAFLD-related questions, revealing high scores for accuracy (mean score, 4.84), completeness (mean score, 2.08 on a 3-point scale), and clarity (mean score, 2.87 on a 3-point scale) [12]. Additionally, ChatGPT has demonstrated proficiency in passing the Iranian medical license test across specialties, excelling in surgery and internal medicine [15]. Furthermore, ChatGPT performs well on quality measures recommended by the AASLD related to cirrhosis and hepatocellular carcinoma as well as previously published questionnaires for physicians and trainees [13].

Nevertheless, studies have acknowledged that ChatGPT's responses may vary due to training data, context, and nuances in language [8,12,14]. For instance, Yeo, *et al*., highlighted ChatGPT's limitations in identifying specific cut-off values in the management of cirrhosis and guideline recommendations for surveillance/screening of HCC, since the model was not able to provide tailored recommendations based on geographic location [13]. In a study determining the accuracy of information provided by ChatGPT regarding liver cancer surveillance and diagnosis, 15 out of 60 answers (25 %) were considered inaccurate, as ChatGPT did not reliably provide accurate information regarding hepatocellular carcinoma surveillance and diagnosis. Wrong answers included contradictory or falsely reassuring statements, which could impact management and patient outcomes. Inaccurate answers provided by ChatGPT may relate to its reliance on open-source information that may not be peer-reviewed or based on high-quality medical evidence [16].

When analyzing answers related to AIH, several omissions or mistakes were noted. In Q8, for instance, some answers did not specify that in clinical practice that imaging modalities beyond ultrasound are rarely used. In this specific question, most chatbots (except for Claude) failed to guide the reader through the options, frequently listing all imaging methods and describing what the general benefits of each in liver disease rather than specifically for AILD.

Furthermore, the experts highlighted the importance of emphasizing certain information or providing additional details crucial to clinical care that were omitted by some chatbots. For instance, in question Q10, none of the chatbots acknowledged that certain drugs potentially should be discontinued in pregnancy such as MMF, which could lead to teratogenicity [17]. In other cases, such as in Q12, some side effects of pharmacologic therapies were described, but without providing sufficient additional information on how to mitigate or prevent these side effects [18].

Copilot had an interesting feature in which it provided references for its response. However, it frequently used non-scientific sources or misleading information. For example, in Q19 Copilot recommend that corticosteroids could be discontinued in two weeks, which is not a reasonable clinical recommendation [19].

For questions specific to PBC, chatbots did not adequately provide information related to diagnostic and prognostic scores (Q2). For example, Claude and Bard did not mention the Paris criteria, a fundamental tool to evaluate response to UDCA treatment in PBC [20]. For a question related to AIH (Q8), there was a vague and general recommendation for imaging methods, as all chatbots, except for Claude, mentioned CT as a diagnostic tool without specifying its limited ability to depict several characteristic features of early PBC or bile duct lesions [21].

When discussing recommendations for liver transplantation, the Mayo Clinic Model for End-Stage Liver Disease (MELD) was only mentioned by Claude and Copilot, and the latter did not specify a reference score that would be considered adequate for recommending transplant. Moreover, in the therapy and clinical decision questions, Bard frequently listed information without going into detail.

Regarding imaging modalities for PSC, we found that only Copilot acknowledged magnetic resonance cholangiopancreatography (MRCP) as the gold standard diagnostic tool for PSC (Q8).

In Q12, Claude listed pharmacological therapies including immunosuppressants in a misleading way. Finally, Copilot failed to address the potential side effects of the recommended therapies, indicating a gap in providing comprehensive information on treatment management for PSC.

Of note was that real-time internet access (Copilot and Bard) did not positively influence the quality of the responses when compared to pre-trained models (ChatGPT and Claude). A possible explanation is that real-time internet data is not particularly curated for quality, and further algorithmic optimization should be performed to achieve higher-quality responses in such models.

Among the main comments from the experts, several key observations were noted. Answers often lacked depth, failing to provide essential information such as detailing the side effects of pharmacologic therapy or listing generic imaging modalities with limited practical application.

Regarding strengths of individual chatbots, Copilot's practice of citing references in each answer was regarded as a positive attribute. Claude was described as presenting more guidance on diagnosis, therapy, and clinical decision inquiries, including emphasizing which radiological method is more frequently used in clinical practice. ChatGPT, while generally well rated by evaluators, occasionally provided information that was not practical, such as listing Obeticholic Acid as a pharmacological option for PSC, despite its being an unapproved therapy for this indication. Lastly, Bard tended to provide less professional information, often listing data without an explanation.

Overall, liver specialists felt that the evaluated chatbots may provide valuable information about AILD. A higher score for the responses was reported among the evaluators from the Americas when compared with their European counterparts, however no clear reasons were found to explain the trend. Among the four included options, Claude had the best overall performance in almost every field evaluated, although all chatbots had limitations that could be improved upon. Potential factors contributing to these results include the constraint of a 300-word limit for each response, the time period of information updates, and issues in providing answers at the level of professional consultants.

Other studies examining the role of chatbots in assessing liver diseases have shown promising results, especially when providing information to patients and caregivers education [13,14]. Despite the potential benefits, these methods may still disseminate inaccurate information, and thus caution is warranted when clinical use is intended. Therefore, while chatbots can serve as valuable resources, healthcare professionals should exercise critical judgment when utilizing them in clinical practice.

5. Conclusions

Among the four chatbots analyzed, Claude exhibited the best overall performance with a mean score of 7.37. Claude also achieved the highest grades in every specific disease assessment, being the AIH, PBC, and PSC mean scores of 7.38, 7.27, and 7.47, respectively. While chatbots demonstrate promise in AILD clinical decision making support, our findings highlight crucial areas for improvement, including refinement in providing specific advice, improving accuracy, and providing focused and up-to-date information. While chatbots show promising in providing related to AILD, particularly for general practitioners without easy access to specialists, addressing these shortcomings would be essential for enhancing the clinical utility of chatbots in the management of AILDs to ensure their effectiveness as clinical decision-support tools.

Author contributions

Author contributions: Study conception and design: Jimmy Daza, Andreas Teufel; Data collection: Jimmy Daza, Lucas Soares Bezerra, Laura Santamaria, Roberto Rueba-Esteban; Analysis and interpretation of results: Jimmy Daza, Lucas Soares Bezerra, Laura Santamaria, Roberto Rueba-Esteban, Heike Bantel, Marcos Girala, Matthias P Ebert, Florian Van Bömmel, Andreas Geier, Andres Gomez-Aldana, Kevin Yau, Mario Alvares-da-Silva, Markus Peck-Radosavljevic, Ezequiel Ridruejo, Arndt Weinmann, Andreas Teufel; Draft manuscript preparation: Jimmy Daza, Lucas Soares Bezerra, Andreas Teufel; Manuscript revision in English: Kevin Yau. All authors reviewed the results and approved the final version of the manuscript.

Conflicts of interest

None.

Funding

AT received grants from the Sino-German Center for Research Promotion (grant numbers: GZ-1546 and C-0012), the State Ministry of Baden-Wuerttemberg for Sciences, Research and Arts supporting the Clinical Cooperation Unit Healthy Metabolism at the Center for Preventive Medicine and Digital Health (grant identifier: CCU Healthy Metabolism), the Baden-Wuerttemberg Center for Digital Early Disease Detection and Prevention (grant identifier: BW- ZDFP), and the Federal Ministry of Education and Research (BMBF) (grant identifier: Q-HCC, LeMeDaRT).

References

[1] D'Amato D, Carbone M. Prognostic models and autoimmune liver diseases. Best Pract Res Clin Gastroenterol 2023;67:101878. https://doi.org/10.1016/j.bpg.2023.101878.
[2] Scaravaglio M, Carbone M, Invernizzi P. Autoimmune liver diseases. Minerva Gastroenterol 2023;69:7–9. https://doi.org/10.23736/S2724-5985.22.03279-X.
[3] Pellicano R, Ferro A, Cicerchia F, Mattivi S, Fagoonee S, Durazzo M. Autoimmune hepatitis and fibrosis. J Clin Med 2023;12:1979. https://doi.org/10.3390/jcm12051979.
[4] Trivella J, John BV, Levy C. Primary biliary cholangitis: epidemiology, prognosis, and treatment. Hepatol Commun 2023;7:e0179. https://doi.org/10.1097/HC9.0000000000000179.
[5] Rawla P, Samant H. Primary sclerosing cholangitis. Statpearls, treasure island (FL). StatPearls Publishing; 2024.
[6] Czaja AJ. Difficult treatment decisions in autoimmune hepatitis. WJG 2010;16:934. https://doi.org/10.3748/wjg.v16.i8.934.
[7] Rahim MN, Liberal R, Miquel R, Heaton ND, Heneghan MA. Acute severe autoimmune hepatitis: corticosteroids or liver transplantation? Liver Transpl 2019;25:946–59. https://doi.org/10.1002/lt.25451.
[8] 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;6:1169595. https://doi.org/10.3389/frai.2023.1169595.
[9] Tang J, James L, Howell M, Tong A, Wong G. eHealth interventions for solid organ transplant recipients: a systematic review and meta-analysis of randomized controlled trials. Transplantation 2020;104:e224–35. https://doi.org/10.1097/TP.0000000000003294.
[10] Holderried M, Hoeper A, Holderried F, Heyne N, Nadalin S, Unger O, et al. Attitude and potential benefits of modern information and communication technology use and telemedicine in cross-sectoral solid organ transplant care. Sci Rep 2021;11:9037. https://doi.org/10.1038/s41598-021-88447-6.
[11] Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci 2023;39:605–7. https://doi.org/10.12669/pjms.39.2.7653.
[12] Pugliese N, Wai-Sun Wong V, Schattenberg JM, Romero-Gomez M, Sebastiani G, Expert Chatbot Working Group NAFLD, et al. Accuracy, reliability, and comprehensibility of chatGPT-generated medical responses for patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2023 S1542-3565(23)00704-8. https://doi.org/10.1016/j.cgh.2023.08.033.
[13] Yeo YH, Samaan JS, Ng WH, Ting P-S, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29:721–32. https://doi.org/10.3350/cmh.2023.0089.
[14] Endo Y, Sasaki K, Moazzam Z, Lima HA, Schenk A, Limkemann A, et al. Quality of chatGPT responses to questions related to liver transplantation. J Gastrointest Surg 2023;27:1716–9. https://doi.org/10.1007/s11605-023-05714-9.
[15] Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. BMJ Health Care Inform 2023;30:e100815. https://doi.org/10.1136/bmjhci-2023-100815.
[16] Cao JJ, Kwon DH, Ghaziani TT, Kwo P, Tse G, Kesselman A, et al. Accuracy of information provided by chatGPT regarding liver cancer surveillance and diagnosis. AJR Am J Roentgenol 2023;221:556–9. https://doi.org/10.2214/AJR.23.29493.
[17] Si T, Huang Z, Hegarty R, Ma Y, Heneghan MA. Systematic review with meta-analysis: outcomes of pregnancy in patients with autoimmune hepatitis. Aliment Pharmacol Ther 2022;55:1368–78. https://doi.org/10.1111/apt.16924.
[18] Coscia LA, Armenti DP, King RW, Sifontis NM, Constantinescu S, Moritz MJ. Update on the teratogenicity of maternal mycophenolate mofetil. J Pediatr Genet 2015;4:42–55. https://doi.org/10.1055/s-0035-1556743.
[19] Terziroli Beretta-Piccoli B, Mieli-Vergani G, Vergani D. Autoimmune hepatitis: standard treatment and systematic review of alternative treatments. World J Gastroenterol 2017;23:6030–48. https://doi.org/10.3748/wjg.v23.i33.6030.
[20] Czaja AJ. Diagnosis and management of the overlap syndromes of autoimmune hepatitis. Can J Gastroenterol 2013;27:417–23. https://doi.org/10.1155/2013/198070.
[21] Zhang Y, Zheng T, Huang Z, Song B. CT and MR imaging of primary biliary cholangitis: a pictorial review. Insights Imaging 2023;14:180. https://doi.org/10.1186/s13244-023-01517-3.