ORIGINAL ARTICLE

Quality of AI chatbot-generated information on hypersensitivity pneumonitis for clinical and patient use

DERYA YENIBERTIZ[1], GÜZIDE TOMAS[2]

[1]Ankara Oncology Training and Research Hospital, University of Health Sciences, Department of Chest Diseases, Ankara; [2]İstanbul Sultan Abdülhamid Han training and research hospital, Chest Disease Clinic, İstanbul, Türkiye

ABSTRACT

Background and Aim: Hypersensitivity pneumonitis (HP) is a complex, immüne mediated interstitial lung disease in which accurate diagnosis and long term management require integration of clinical, radiologic, and exposure-related information. Patients increasingly use artificial intelligence (AI) based chatbots to obtain disease related information; however, the quality, readability, and patient usability of such content remain unclear. This study aimed to evaluate the quality, reliability, readability, and patient-centered usability of AI chatbot generated information on HP.

Materials and Methods: Using Google Trends, we identified four of the most frequently searched patient-oriented questions regarding HP: (1) What is HP and what causes it? (2) What are the clinical features of HP? (3) How is HP treated? (4) How is HP diagnosed? These questions were submitted verbatim to eight AI chatbots (ChatGPT-5.1, Claude 3, Microsoft Copilot, DeepSeek V3, Gemini Pro, Grok 4, Kimi K2, Perplexity AI). A total of 32 responses were independently evaluated in a blinded fashion by four pulmonology professors specializing in interstitial lung diseases. Content quality and reliability were assessed using DISCERN; understandability and actionability with PEMAT-P; global written readability with the Written Readability Rating (WRR); and structural readability with the Flesch–Kincaid Grade Level (FKGL).

Results: All chatbot outputs required advanced literacy, with FKGL scores ranging from 20.17 to 29.07 and a mean of approximately 24–25, indicating college or postgraduate reading level. No chatbot produced content within the recommended patient-appropriate range (FKGL ≤ 8). WRR scores declined with increasing clinical complexity, from 67.85 for definitional content (Q1) to 51.227 for diagnostic explanations (Q4). DISCERN scores varied substantially across models (35.001–57.103), with most chatbots falling into the “fair–good” range, reflecting partially reliable but incomplete information. PEMAT-P understandability scores were moderate (highest: 66.302% for DeepSeek V3), whereas actionability was consistently low, indicating a lack of clear, concrete patient instructions.

Received: 15 December 2025 | Accepted: 4 March 2026

Corresponding Author: Derya YENIBERTIZ / Department of Pulmonology, University of Health Sciences, Dr. Abdurrahman Yurtaslan Ankara Oncology Training and Research Hospital; Ankara, Türkiye / E-mail: yenibertizderya@gmail.com
ORCID: 0000-0002-1783-4015

Conclusion: AI chatbots can generate clinically rich explanations of HP but currently produce content that is too complex and insufficiently actionable for most patients. The observed "readability–quality gap" suggests that chatbot-generated information should not yet replace professionally curated patient education materials. A hybrid model in which AI-generated drafts are refined by clinical experts may represent the most appropriate strategy for safe integration of AI tools into patient education on complex interstitial lung diseases such as HP.

Key words: Hypersensitivity pneumonitis, artificial intelligence, chatbots.

Introduction

Hypersensitivity pneumonitis (HP) is a complex, immune-mediated interstitial lung disease resulting from repeated inhalational exposure to a wide variety of organic and inorganic antigens. The disease encompasses acute, subacute, and chronic phenotypes, each influenced by antigen dose, exposure frequency, host susceptibility, and immunologic reactivity(1). Recent consensus classifications have further emphasized the distinction between fibrotic and non-fibrotic HP, as fibrosis represents the strongest predictor of disease progression and mortality(1, 2). Despite advances in diagnostic imaging and exposure assessment, HP remains one of the most challenging interstitial lung diseases (ILDs) to diagnose due to its broad clinical heterogeneity, overlapping features with other ILDs, and the absence of a single definitive diagnostic marker(3). Epidemiological data indicate that HP is underdiagnosed and frequently misclassified, especially in its chronic fibrotic form, which can mimic idiopathic pulmonary fibrosis both radiologically and clinically (4). Exposure history remains central to diagnosis, yet patients are often unaware of environmental or occupational antigen sources, making accurate identification difficult without structured clinician-guided assessment. Imaging characteristics, including mosaic attenuation, air-trapping, centrilobular nodules, and the "three-density pattern," provide valuable clues; however, interpretation requires specialized radiologic expertise, and variability across institutions further complicates diagnostic pathways(3). In this context of clinical uncertainty and diagnostic complexity, patients increasingly turn to online health information to understand their symptoms, potential exposures, diagnostic tests, and treatment options. Studies have shown that up to 70% of individuals with chronic diseases search the internet before or after medical visits, particularly for rare or poorly understood conditions (5). However, the quality and readability of web-based medical content vary widely. Research evaluating online ILD-related information has demonstrated significant deficiencies in accuracy, completeness, and readability, with most materials written far above recommended health literacy levels(6). This discrepancy is concerning because limited health literacy is independently associated with poorer disease understanding, reduced adherence, and worse health outcomes. Meanwhile, AI powered chatbots and large language models (LLMs) have emerged as prominent tools for generating on-demand medical explanations. Their ability to synthesize large bodies of biomedical text has accelerated their adoption in clinical support, patient engagement, and public health communication. Early evidence suggests that AI chatbots can provide medically relevant, cohesive explanations; however, studies also reveal substantial variation in accuracy, depth, and readability—particularly when addressing nuanced or complex clinical topics(7). For diseases such as HP, which require precise environmental, imaging, and immunologic interpretation, the reliability and patient-friendliness of AI-generated responses remain insufficiently characterized. To ensure alignment with real-world informational needs, the present study selected HP-related questions using Google Trends, thereby capturing the most frequently searched patient-oriented queries on hypersensitivity pneumonitis. This method provides an ecologically valid representation of public concerns and information

gaps, allowing evaluation of AI performance on the exact topics patients search most frequently online. Accordingly, this study aims to conduct a descriptive, comparative content analysis of responses generated by eight widely used AI chatbots. Using validated instruments DISCERN, PEMAT-P, WRR, and FKGL we evaluated the quality, reliability, comprehensibility, and readability of chatbot-generated information. By systematically analyzing AI-produced answers to patient-centered questions about HP, this work seeks to elucidate the strengths and limitations of current AI tools, inform their responsible integration into patient education, and identify key areas for improvement in future healthcare-oriented LLM development.

Material and method

Study design and question selection

Query selection was conducted systematically to ensure that the evaluated questions accurately reflected real-world patient information-seeking behavior while maintaining clinical relevance. To identify the most frequently searched patient-oriented questions related to HP, Google Trends data were analyzed over a defined 12-month period (January 2023 to January 2024) using global search parameters. Google Trends was selected as it captures large-scale, real-time public search behavior and provides an ecologically valid representation of the health-related topics most commonly sought by patients outside the clinical setting. This approach has been widely used in prior studies to approximate authentic patient concerns in the absence of direct patient interviews. Based on relative search interest and clinical relevance, four core questions were selected for detailed evaluation: (1) *What is hypersensitivity pneumonitis and what causes it?* (2) *What are the clinical features of hypersensitivity pneumonitis?* (3) *How is hypersensitivity pneumonitis treated?* and (4) *How is hypersensitivity pneumonitis diagnosed?* These questions were intentionally chosen to represent a gradient of informational complexity, ranging from basic disease definition to diagnostic reasoning. All questions were used exactly as identified, without modification or simplification, and were presented identically to all evaluated AI chatbots to ensure standardization and comparability across generated responses.

Expert evaluators

Four professors specializing in interstitial lung diseases served as expert evaluators. Each expert independently reviewed and scored all chatbot responses in a blinded manner. The evaluators were unaware of which chatbot produced each response. Evaluation Metrics, Chatbot-generated texts were assessed in four main domains: Content Quality and Reliability; Medical accuracy, completeness, scientific validity, and evidence-based information were assessed using the DISCERN instrument: The DISCERN tool consists of 16 items. Each item is scored from 1 to 5 (higher scores indicate higher quality). It evaluates three core dimensions: Reliability of the publication, quality of treatment information, overall quality rating. Understandability of Written Health Information; Two validated international tools were used to assess how understandable and patient-friendly the texts were: PEMAT-P (Patient Education Materials Assessment Tool – Printable); A 17-item tool assessing *understandability* and *actionability*. Scored as yes/no/not applicable. A percentage score is calculated. WRR (Written Readability Rating); A global 1–5 rating assessing the clarity and simplicity of the writing. Readability Analysis; Structural readability of the texts was assessed using the Flesch–Kincaid Grade Level (FKGL) formula, widely used for evaluating reading difficulty:

$$FKGL = 0.39 \times \left(\frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \times \left(\frac{\text{total syllables}}{\text{total words}} \right) - 15.59$$

For each chatbot response, the following were calculated using a Python-based linguistic analysis script: total word count, total sentence count, total syllable count. Data Collection and Analysis Procedure; All chatbot responses were collected manually and stored in structured text files. Expert reviewers independently scored each response

using DISCERN, WRR, and PEMAT-P. FKGL scores were calculated automatically using Python. Aggregated data were analyzed descriptively, including: means and medians, variation coefficients, inter-chatbot comparison. The performance of chatbots was compared with established health literacy recommendations.

Data collection and statistical analysis

Aggregated data were summarized using descriptive statistics. Continuous variables are presented as means ± standard deviations (SD) when normally distributed and as medians with interquartile ranges (IQR) when non-normally distributed. For chatbot-level comparisons, mean scores were used to summarize overall performance across evaluators, as reflected in Tables 1–3. Variability across chatbot outputs was assessed using standard deviation and range values. All statistical summaries were generated following inspection of score distributions.

Ethical considerations

No human subjects were involved, and expert reviewers only provided professional scoring; therefore, formal ethics committee approval was not required. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Results

In this study, eight AI based chatbots were evaluated for the accuracy, readability, and patient-centered usability of their responses to four frequently searched questions regarding HP. A total of 32 chatbot-generated outputs were assessed independently by four pulmonology experts using the DISCERN, PEMAT-P, WRR, and FKGL tools. Descriptive results are presented as means and, where appropriate, complemented by measures of variability (e.g., standard deviation), as summarized in Tables 1–3.

Inter-rater reliability (ICC)

To assess the robustness and reproducibility of expert-based scoring, inter-rater reliability among the four pulmonology evaluators was examined using intraclass correlation coefficients (ICC; two-way random effects model, absolute agreement). For DISCERN total scores, the single-measure ICC demonstrated moderate agreement (ICC (2,1) =0.44), while the average-measures ICC based on the mean of four raters indicated good reliability (ICC (2,4) =0.76). For PEMAT-P understandability scores, the single-measure ICC showed low agreement (ICC (2,1) =0.23), which improved to moderate reliability when averaged across raters (ICC (2,4) =0.54). These findings support the methodological robustness of using mean expert ratings in the comparative evaluation of chatbot performance. Exploratory Inferential Analysis (Kruskal–Wallis); In response to reviewer recommendations, exploratory between-chatbot comparisons were performed using the non-parametric Kruskal–Wallis test. Statistically significant differences were observed between chatbots in overall information quality (DISCERN scores; p=0.030). In contrast, chatbot differences in PEMAT-P understandability did not reach statistical significance (p=0.142), suggesting broadly similar limitations in patient-oriented clarity across models despite variability in overall content quality.

Readability analysis – FKGL

The Flesch–Kincaid Grade Level (FKGL) scores demonstrated that all chatbot responses

Table 1. Flesch–Kincaid Grade Level (FKGL) Scores for All Chatbot Responses

Chatbot	FKGL Score
Chatbot 1	27.09
Chatbot 2	26.68
Chatbot 3	23.72
Chatbot 4	20.17
Chatbot 5	23.26
Chatbot 6	25.04
Chatbot 7	29.07
Chatbot 8 (Perplexity AI)	**22.20**

Notes: FKGL: values represent mean grade-level scores calculated from chatbot-generated responses using the Flesch–Kincaid formula. Higher scores indicate increased reading difficulty. All values exceeded grade level 20, indicating university- or postgraduate-level readability.

required advanced literacy levels, substantially exceeding the recommended 5th–8th grade readability range for patient education materials. As shown in Table 1, FKGL values ranged from 20.17 to 29.07, indicating that content generally required college-level or postgraduate-level reading ability. The lowest FKGL scores were observed in DeepSeek V3 (20.17) and Perplexity (22.20), suggesting relatively simpler but still challenging language. In contrast, Kimi K2 (29.07), ChatGPT-5.1 (27.09), and Claude 3 (26.68) produced the most complex and academically styled responses. Overall, the mean FKGL score was approximately 24–25, and no chatbot produced material within the ideal patient-appropriate readability range (FKGL ≤ 8). The extensive use of HP-specific terminology including fibrosis, BAL, precipitins, and HRCT pattern descriptions likely contributed to elevated scores.

Table 2. Mean Writing Readability Rating (WRR) by Question

Question No.	Question Content	Mean WRR
Q1	"What is HP and what are its causes?"	**67.85**
Q2	"How does HP present clinically?"	**58.125**
Q3	"How is HP treated?"	**57.928**
Q4	"How is HP diagnosed?"	**51.227**

Notes: WRR scores declined as question complexity increased. The highest readability was observed for definition-based content (Q1), while diagnostic explanations (Q4) demonstrated the lowest readability.

Writing Readability Rating (WRR)

WRR scores exhibited a clear downward trend across the four evaluated questions (Table 2). The highest readability was found in responses to definitional content Q1: "What is HP and what are its causes?" (WRR = 67.85), suggesting that chatbots conveyed basic disease concepts more effectively. Conversely, diagnostic explanations demonstrated the lowest readability (Q4 WRR = 51.227). This is consistent with clinical practice, where diagnosing chronic HP may require multidisciplinary integration and, in selected cases, invasive procedures such as transbronchial biopsy, adding further complexity to patient-facing explanations(8). The reduction in readability appears to be driven by: complex, multistep diagnostic pathways, dense clinical terminology, explanations structured for clinicians rather than patients.

The observed gradient Definition > Symptoms > Treatment > Diagnosis suggests that as clinical complexity increased, chatbot-generated text became less accessible.

DISCERN assessment

Expert evaluation of response quality revealed that chatbots delivered adequate information regarding disease definitions and general explanations. However, several limitations were noted, including insufficient discussion of:

Table 3. DISCERN and PEMAT-P Scores of Chatbot Responses (Mean ± SD)

Chatbot	DISCERN (Mean ± SD)	PEMAT Understandability % (Mean ± SD)	PEMAT Actionability % (Mean ± SD)
ChatGPT-5.1	44.06 ± 8.16	70.50 ± 3.00	43.00 ± 11.43
Claude 3	57.11 ± 7.67	65.75 ± 6.50	39.50 ± 7.00
Microsoft Copilot	41.25 ± 1.65	62.75 ± 8.96	35.75 ± 18.46
DeepSeek V3	53.05 ± 10.12	67.25 ± 11.84	64.00 ± 8.08
Gemini Pro	51.25 ± 6.35	70.25 ± 10.69	43.00 ± 19.80
Grok 4	35.00 ± 5.35	51.75 ± 8.02	21.50 ± 18.59
Kimi K2	49.22 ± 3.87	64.25 ± 9.50	43.00 ± 16.17
Perplexity AI	51.09 ± 10.21	60.75 ± 9.50	25.00 ± 21.56

Notes: DISCERN and PEMAT-P scores are reported as mean ± standard deviation (SD) across four expert evaluators. Higher DISCERN scores indicate higher information quality. PEMAT-P scores are expressed as percentages; higher percentages indicate better understandability/actionability.

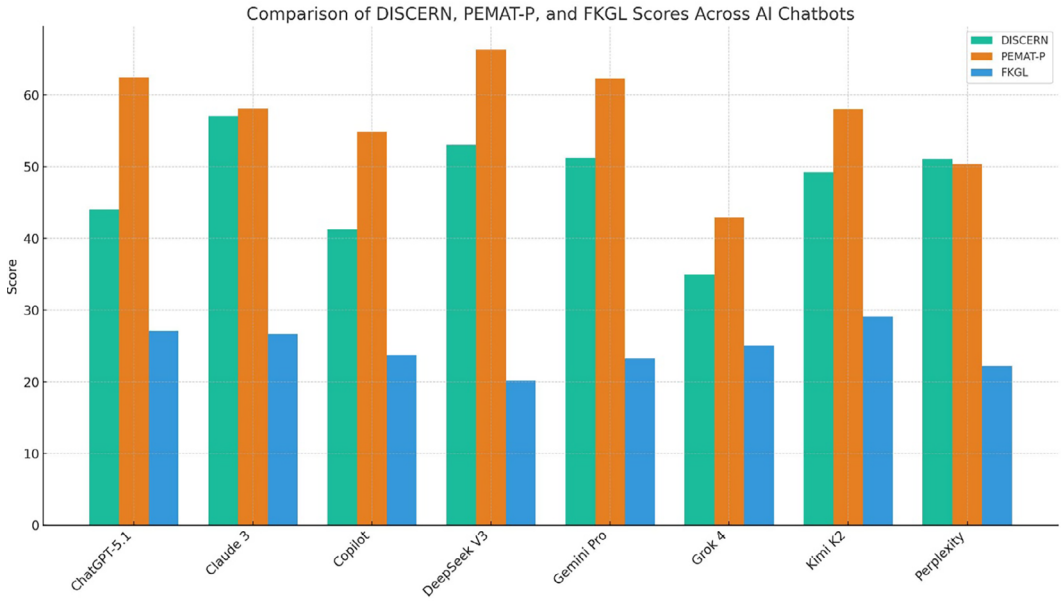

Figure 1. Comparison of DISCERN, FKGL and PEMAT-P Scores Across Eight AI Chatbots Evaluating Hypersensitivity Pneumonitis Information.

- treatment risks and benefits,
- uncertainties and prognosis,
- alternative therapeutic options.

As shown in Table 3, DISCERN scores ranged widely. Claude 3 achieved the highest quality rating (57.103), while Grok 4 scored lowest (35.001), indicating considerable variability in reliability and balance of medical content. Many chatbots scored within the "fair–good" DISCERN range (40–60), reflecting partially reliable but incomplete medical information.

PEMAT-P (Understandability and Actionability)

PEMAT-P results showed moderate understandability across the eight chatbots (Table 3). Strengths included clear subheadings, logical organization, and partial simplification of complex concepts. The highest understandability score was achieved by DeepSeek V3 (66.302%), nearing the commonly recommended threshold for patient educational materials (≥70%).

In contrast, actionability was consistently low across all models. Chatbots often failed to:

- provide concrete instructions regarding environmental antigen avoidance,
- outline step-by-step patient-centered actions,
- translate treatment principles into practical guidance.

Given that HP management relies heavily on environmental control and exposure mitigation, low actionability poses a critical limitation for patient usability.

Question-based performance comparison

Across all evaluated metrics (FKGL, WRR, DISCERN, PEMAT-P), a consistent pattern emerged:

- The highest-performing question was Q1 (definition of HP),
- While the lowest-performing was Q4 (diagnosis of HP).

This pattern indicates that chatbots communicate introductory biomedical concepts more effectively than complex diagnostic information. Diagnostic reasoning, interpretation of imaging findings, and decision-making algorithms appear to challenge the generative models' ability to maintain patient-friendly communication.

Integrated interpretation of findings

The findings of this study show that:

- Chatbots demonstrate strong informational depth.
  They provide detailed and clinically correct descriptions of HP, reflecting robust underlying biomedical knowledge.
- Readability remains a major barrier.
  All FKGL scores fall within postgraduate literacy levels.
- User-friendliness is limited.
  WRR scores decrease substantially as content complexity increases.
- Action-oriented guidance is weak.
  Low PEMAT-P actionability underscores limited usefulness for patient self-management.
- Chatbot outputs are clinically appropriate but overly complex for patients.
  Responses resemble clinician-level explanations rather than patient education materials.
- Meaningful variability exists among models.
  Differences in DISCERN, PEMAT-P, FKGL, and WRR scores reflect heterogeneous linguistic and structural performance.

Discussion

Our evaluation of eight AI chatbots' responses to frequently asked questions about HP demonstrates a consistent pattern: while AI-generated outputs often contain substantial clinical content, they remain largely inaccessible for patients due to high readability demands, and frequently lack actionable guidance or robust referencing. These findings align with and extend emerging evidence from recent evaluations of AI based health communication tools. Several recent studies have highlighted similar limitations of AI chatbots and generative models when used to produce patient facing medical information. In a recent cross-sectional analysis of AI-generated patient education for anterior cruciate ligament injury and treatment, authors found that although the overall medical information quality (assessed by DISCERN) was moderate, none of the evaluated models achieved acceptable readability for general public use(9). Another study comparing readability and reliability of patient information leaflets generated by AI models (on topics including Alzheimer's, dementia, and delirium) similarly reported that model outputs significantly exceeded recommended readability thresholds, and that there were meaningful readability differences among models(10). A systematic review addressing the broader role of AI in health literacy emphasized that although AI techniques are increasingly adopted to produce educational materials, only a small minority meet accepted readability standards; moreover, few studies engage real patients or focus on actionability in recommendations(11). In a clinical-query context, a study testing responses from several AI chatbots on urology questions revealed that while factual accuracy was frequently acceptable, there was substantial variation in completeness and reliability, particularly for complex or less common conditions(12). Finally, a large cross-specialty evaluation assessing the accuracy and comprehensiveness of chatbot responses to physician-generated medical queries reported that although many answers were largely correct, completeness was often suboptimal especially regarding nuances, contraindications, and guideline-specific caveats(7). Our findings echo and reinforce these observations especially regarding readability: our FKGL analysis showed all chatbot outputs required college-level literacy, far above the 6th–8th grade level widely recommended for patient education materials. WRR scores declined sharply for diagnostic content, and PEMAT-P actionability remained low across all models.

Implications for patient education and clinical use; given the complexity of HP a disease that often requires interpretation of environmental exposures, imaging, immunological testing, and long-term management the limitations documented here carry significant practical consequences:

1. Accessibility gap: High reading level and limited actionability likely make AI-generated HP information inadequate for many patients, especially those with limited health literacy. This may contribute to misunderstanding of diagnosis, delayed care-seeking, or mismanagement of environmental exposures.
2. Risk of misinterpretation: Without proper referencing (as indicated by low DISCERN scores), patients may accept confident-sounding but incomplete or non-evidence-based recommendations as authoritative. AI chatbots may thus inadvertently propagate misinformation or oversimplified guidance.
3. Uneven model performance: The observed heterogeneity between chatbots suggests that some models (e.g., those with relatively better readability or quality scores) may perform better in certain contexts, but none currently meet the dual criteria of high reliability *and* patient-level accessibility.
4. Need for supervised or hybrid deployment: Given limitations, chatbots may have utility as support tools for clinicians, patient education draft generation, or pre-visit question preparation but should not yet replace professional consultation or validated patient materials.

Use of multiple validated assessment instruments (DISCERN, PEMAT-P, WRR, FKGL) provides a multidimensional evaluation of the quality, readability, and usability of chatbot-generated information. The inclusion of several widely used chatbots offers a broad representation of current AI tools, while the focus on hypersensitivity pneumonitis (HP) a complex interstitial lung disease creates a challenging and clinically relevant test case for assessing the suitability of AI-generated patient education materials. Based on our findings and the context of existing literature, several recommendations emerge for advancing the safe and effective use of AI in patient communication. First, a hybrid human AI workflow should be adopted. AI may be used to generate initial drafts of patient-facing materials, but expert clinical review remains essential to simplify language, ensure evidence-based accuracy, and incorporate practical, actionable guidance. Second, model fine tuning for readability and usability is needed. Developers should prioritize not only factual correctness but also health literacy appropriate language, clear and coherent structure, minimization of unnecessary medical jargon, and transparent citation of information sources. Third, user-centered evaluation is critical. Future research should directly involve patients—particularly those with limited health literacy—to assess comprehension, retention, and real-world usability of AI-generated health texts. Fourth, professional guidelines are needed to define minimum acceptable standards for readability, referencing, and actionability before AI-derived content can be safely used in patient education. Finally, broader studies should assess AI performance across a wider range of diseases and over longer periods to evaluate risks such as misinformation drift and the need for systematic updates as clinical guidelines evolve.

Limitations

This study evaluated standardized, single-turn, English-language responses generated by eight AI chatbots to four patient-oriented questions identified through Google Trends. While this design ensured comparability across models, real-world chatbot use often involves iterative prompting, follow-up clarification, and personalization, which may influence response quality and accessibility. Inter-rater reliability analyses supported the robustness of the expert-based evaluation, particularly when scores were averaged across reviewers, consistent with our reporting approach. Although validated proxy tools (DISCERN and PEMAT-P) were used to assess quality and usability, patient comprehension and perceived usefulness were not directly measured. Finally, this analysis focused on overall quality, readability, and usability metrics rather than exhaustive guideline-level verification of each clinical statement; future studies incorporating patient co-design and multi-turn interactions may further strengthen generalizability.

Conclusion

In summary, while AI chatbots show considerable promise for generating comprehensive and clinically detailed health information, their current outputs for complex diseases like HP remain difficult for the average patient to understand and apply. Our findings—consistent with growing evidence in the literature—highlight a persistent "readability–quality tradeoff." Until AI-generated content is optimized for clarity, reliability, and patient usability, these tools should be employed cautiously and under professional supervision. A hybrid workflow that integrates AI-generated drafts with expert clinical refinement appears to be the most effective and safe strategy for leveraging AI's strengths while mitigating its limitations.

Ethics approval and consent to participate: This study did not involve human participants, patient data, or animals. The evaluated material consisted solely of publicly accessible outputs generated by artificial intelligence chatbots and anonymized expert ratings. In accordance with national regulations and institutional policies, formal ethics committee approval was not required. The study was conducted in line with the principles of the Declaration of Helsinki.

Consent for publication: Not applicable. This manuscript does not contain any individual person's data in any form (including individual details, images, or videos).

Availability of data and materials: The datasets generated and analyzed during the current study (chatbot responses and anonymized scoring sheets) are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions: D.Y. and G.T. conceived and designed the study. D.Y. collected the chatbot outputs and performed the primary data organization. G.T. coordinated expert evaluations and contributed to the interpretation of the findings. Both authors contributed to drafting and critically revising the manuscript, approved the final version, and agree to be accountable for all aspects of the work.

Acknowledgements: The authors would like to thank the pulmonology professors who served as expert evaluators for their valuable time and contributions to the scoring of chatbot-generated content.

References

1. Calaras D, David A, Vasarmidi E, Antoniou K, Corlateanu A. Hypersensitivity Pneumonitis: Challenges of a Complex Disease. Can Respir J. 2024 Jan 18;2024:4919951. doi: 10.1155/2024/4919951.
2. Hamblin M, Prosch H, Vašáková M. Diagnosis, course and management of hypersensitivity pneumonitis. European Respiratory Review. 2022;31(163).
3. Dabiri M, Jehangir M, Khoshpouri P, Chalian H. Hypersensitivity Pneumonitis: A Pictorial Review Based on the New ATS/JRS/ALAT Clinical Practice Guideline for Radiologists and Pulmonologists. Diagnostics (Basel). 2022 Nov 20;12(11):2874. doi: 10.3390/diagnostics12112874.
4. Magee AL, Montner SM, Husain A, Adegunsoye A, Vij R, Chung JH. Imaging of Hypersensitivity Pneumonitis. Radiol Clin North Am. 2016 Nov;54(6):1033-1046. doi: 10.1016/j.rcl.2016.05.013.
5. Zhao YC, Zhao M, Song S. Online health information seeking among patients with chronic conditions: integrating the health belief model and social support theory. Journal of medical Internet research. 2022;24(11):e42447.
6. Daraz L, Morrow AS, Ponce OJ, Farah W, Katabi A, Majzoub A, et al. Readability of Online Health Information: A Meta-Narrative Systematic Review. Am J Med Qual. 2018 Sep/Oct;33(5):487-492. doi: 10.1177/1062860617751639.
7. Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 2023 Oct 2;6(10):e2336483. doi: 10.1001/jamanetworkopen.2023.36483.
8. Bezerra Botelho A, Ferreira RG, Coletta ENAM, Cerezoli MT, Martins RB, Gomes PS, et al. Transbronchial biopsy in chronic hypersensitivity pneumonitis . Sarcoidosis Vasc Diffuse Lung Dis [Internet]. 2021 Jun. 28 [cited 2026 May 21];38(2):e2021018. Available from: https://

mattioli1885journals.com/index.php/sarcoidosis/article/view/8998
9. Fahy S, Oehme S, Milinkovic D, Jung T, Bartek B. Assessment of Quality and Readability of Information Provided by ChatGPT in Relation to Anterior Cruciate Ligament Injury. J Pers Med. 2024 Jan 18;14(1):104. doi: 10.3390/jpm14010104.
10. Jido JT, Al-Wizni A, Le Aung S. Readability of AI-generated patient information leaflets on Alzheimer's, vascular dementia, and delirium. Cureus. 2025;17(6).
11. Abeo ANA, Armstrong S, Scriney M, Goss H. Artificial Intelligence Techniques and Health Literacy: A Systematic Review. Mayo Clin Proc Digit Health. 2025 Sep 24;3(4):100269. doi: 10.1016/j.mcpdig.2025.100269.
12. Carlson JA, Cheng RZ, Lange A, Nagalakshmi N, Rabets J, Shah T, et al. Accuracy and Readability of Artificial Intelligence Chatbot Responses to Vasectomy-Related Questions: Public Beware. Cureus. 2024 Aug 28;16(8):e67996. doi: 10.7759/cureus.67996.

Copyright: The Author(s), 2026. Licensee Mattioli 1885, Fidenza, Italy. This is an open-access article distributed under the terms of the Creative Commons Attribution NonCommercial License (CC BY-NC-4.0).
Disclaimer/Publisher's Note: The statements, opinions and data contained in this article are solely those of the author(s) and contributor(s) and do not necessarily reflect those of their affiliated organizations, the publisher, the editors or the reviewers. The publisher and the editors disclaim any responsibility for injury to people or property resulting from any ideas, methods, instructions or products mentioned in the content. Any product that may be evaluated in this article, or claim made by its manufacturer, is not guaranteed or endorsed by the publisher.