Respiratory Research

RESEARCH Open Access

Accuracy, comprehensiveness and understandability of AI-generated answers to questions from people with COPD: the AIR-COPD Study

Mattia Nigro[1,2], Greta E. Behring[1,2], Andrea Aliverti[3], Alessandra Angelucci[3], Anita Kay Simonds[4,5], Antonio Anzueto[6], Peter Martin Calverley[7], Francesco Amati[1,2], Anna Stainer[1,2], Apostolos Bossios[8,9,10], Hilary Pinnock[11], Jeanette Boyd[12], Pippa Powell[12], Stefano Aliberti[1,2*] and AIR-COPD Task Force

Abstract

Background Chronic obstructive pulmonary disease (COPD) remains an underestimated and underdiagnosed condition due to low disease awareness. Generative Artificial Intelligence (AI) chatbots are convenient and accessible sources of medical information, but evaluation of the quality of answers provided by patient-generated questions about COPD has not been performed to date.

Objective To assess and compare accuracy, comprehensiveness, understandability and reliability of different AI chatbots in response to patient-generated questions on the clinical management of COPD.

Methods A cross-sectional study was conducted in collaboration with the European Respiratory Society (ERS), the European Lung Foundation (ELF), and the ERS CONNECT Clinical Research Collaboration (CRC). Fifteen real questions formulated by ELF COPD patient representatives were divided into three difficulty tiers (easy, medium, difficult) and submitted to ChatGPT (version 3.5), Bard, and Copilot. Experts assessed accuracy and comprehensiveness on a 0–10 scale; patients assessed understandability using the same scale. Reliability was assessed by two investigators. Reviewers were blinded to which AI system generated the answers, and only those who completed all evaluations were included in the analysis.

Results ChatGPT responses were the most reliable (14/15), followed by Copilot (12/15) and Bard (11/15). ChatGPT scored higher for accuracy (8.0 [7.0 – 9.0]) and comprehensiveness (8.0 [6.8 – 9.0]) than Bard (6.0 [5.0 – 8.0] and 6.0 [5.0 – 7.0]) and Copilot (6.0 [5.0 – 7.3] and 6.0 [5.0 – 8.0]) (both $P<0.001$). Understandability was similar across all software (ChatGPT: 8.0 [8.0–10.0]; Bard: 9.0 [8.0–10.0]; Copilot: 9.0 [8.0–10.0]) ($P=0.53$). No significant effect was detected according to the difficulty of the question.

Conclusion Our findings suggest that AI chatbots, particularly ChatGPT, can provide accurate, comprehensive and understandable answers to patients' questions.

*Correspondence:
Stefano Aliberti
stefano.aliberti@hunimed.eu

Full list of author information is available at the end of the article

Keywords COPD, Artificial intelligence, AI, Reliability, Accuracy, Comprehensiveness, Understandability, Disease awareness

Introduction

Chronic obstructive pulmonary disease (COPD) remains an underestimated and underdiagnosed condition, largely due to limited public and professional awareness, under-recognition of symptoms and signs, and inadequate performance of appropriate functional testing [1]. Even after a correct diagnosis is established, disease management can be challenging because of the heterogeneity and complexity of COPD in terms of etiology, clinical phenotypes and endotypes, and its frequent association with extrapulmonary manifestations and comorbidities [2, 3]. These diagnostic and therapeutic challenges primarily stem from insufficient awareness of the disease, including limited understanding of its causes, symptoms, pathophysiology, natural history, treatment options, and overall impact on patients' lives [4]. People living with COPD urgently feel the need to better understand their condition. Higher disease awareness and self-management interventions have recently been associated with improvements in health-related quality of life and a lower chance of respiratory-related hospital admissions in COPD [5]. Nonetheless, disease awareness remains generally low and poorly studied [6, 7]. This lack of awareness is evident not only in high-income countries, but also in low- and middle-income settings, where limited recognition of COPD contributes to delayed diagnoses, suboptimal treatment, and poorer clinical outcomes [8]. Therefore, patients may increasingly rely on online resources to seek clinical information or supplementary advice.

In recent years, large language models (LLMs) and generative artificial intelligence (AI) chatbots have emerged as accessible and convenient sources of medical information, particularly for individuals living with chronic diseases [9, 10]. Recent studies have demonstrated promising user acceptance of AI chatbots for chronic disease self-management, with positive feedback regarding their perceived helpfulness, ease of use, and user satisfaction [9]. Furthermore, AI chatbots have shown potential in delivering accurate and comprehensive medical information, with the overall quality of answers reported to be relatively high across different question types and levels of difficulty [11–14].

To date, no rigorous assessment has evaluated the quality of chatbot-generated responses to real patient-derived questions on COPD. A recent study attempted to address this issue but was limited by its use of pre-formulated questions and subjective clinician ratings without standardized evaluation tools [15]. Moreover, it did not account for the variability or complexity of actual patient queries, nor did it stratify responses by clinical domain. These limitations highlight the need for a more precise and methodologically robust investigation.

The aim of this study was to fill this gap by systematically evaluating the reliability and quality of AI chatbot responses to authentic questions about COPD management from people living with the condition.

Materials and methods

Study design and study procedures

This was a cross-sectional study to evaluate the performance of LLM-based chatbots in responding to patient-formulated questions about COPD. Three LLM-based chatbots were selected for this analysis, based on their expected widespread use among the general population at the time of study design: ChatGPT (version 3.5, OpenAI), Bard (now Gemini, Google), and Copilot (Microsoft). The study was developed in collaboration with the European Respiratory Society (ERS) Assemblies, the European Lung Foundation (ELF) Patient Advisory Group (PAG) for COPD, and the ERS CONNECT Clinical Research Collaboration (CRC) [16].

In November 2023, ELF patient representatives living with COPD were invited to formulate questions they would have liked to ask their treating physicians. Questions were required to be syntactically simple, limited to a maximum of two short sentences, and to focus on issues considered relevant from the patient perspective. Two investigators (MN and SA) categorized the submitted questions into three equal-size tiers of difficulty (*i.e.*, easy, medium, and difficult) based on their clinical judgment. Concurrently, a panel of international experts in COPD clinical management—the AIR-COPD Task Force—was assembled through outreach to the ERS Assemblies and the ERS CONNECT CRC.

All questions were submitted to each of the three AI tools on the same day (January 30th, 2024) from the same location (Milan, Italy), using two distinct devices connected to different internet networks and IP addresses. Two investigators (MN and SA) independently assessed the reliability of each AI tool, defined as the proportion of responses that maintained a consistent core of information when the same question was submitted simultaneously on the same software across different devices. In cases of disagreement between the two investigators, consensus was reached through a brief discussion.

Following reliability assessment, a random number generator (RNG) sequence was used to select one of the two response sets per software. The selected responses

were exported to a spreadsheet and uploaded to SurveyMonkey.com for blinded evaluation.

The following performance metrics were evaluated:

i) Accuracy, defined as the degree to which the response conveyed correct and clinically appropriate information, as rated by clinical experts on a 0–10 scale (0 = not accurate at all; 10 = completely accurate);
ii) Comprehensiveness, defined as the extent to which the response addressed all relevant aspects of the question, also evaluated by experts on a 0–10 scale (0 = totally incomplete; 10 = fully complete);
iii) Understandability, defined as the clarity and accessibility of the response, as rated by patients based on their lived experience with COPD, on a 0–10 scale (0 = not understandable at all; 10 = completely understandable).

All evaluations were performed under blinded conditions, and neither the experts nor the patient reviewers were informed of which chatbot generated each response. To prevent selection bias, only reviewers who completed all evaluations across all questions and platforms were included in the final analysis.

Table 1 Complete list of the 15 questions, categorized into three levels of difficulty

Easy	
1	Is COPD always caused by smoking or can you get it because of other things?
2	Can I make the diagnosis of COPD through a chest CT scan?
3	What are the different phases/stages a COPD patient may go through?
4	Is there an official/standard method of treating COPD patients?
5	What causes a COPD exacerbation and what exactly is happening when one happens?
Medium	
6	Is the diagnosis of COPD always accurate and correct?
7	Is there a cure for COPD?
8	Is it possible to slow down the deterioration of breathing in COPD?
9	You get COPD, so harm is already done, so does it matter if you continue smoking?
10	Is it always possible to find the cause of an exacerbation for an individual COPD patient?
Difficult	
11	I have COPD, how long do I have to live?
12	What is the best climate to live in for someone with COPD?
13	I am a COPD patient. Am I always supposed to use antibiotics when I have a disease flare-up?
14	I have asthma and I currently smoke. Do I also have COPD?
15	I have been diagnosed with COPD. How many times per year am I supposed to undergo spirometry?

Statistical analysis

Once evaluations were completed, results were exported to an Excel spreadsheet for analysis. The three AI chatbots were compared in terms of reliability, accuracy, comprehensiveness, and understandability. Categorical variables were presented as proportions, and continuous variables as medians with interquartile ranges [IQR]. Due to non-normal data distribution, the Kruskal–Wallis H test was used for group comparisons of continuous variables. Post-hoc pairwise comparisons were conducted using Dunn's test. All statistical analyses and visualizations were performed using RStudio (version 2023.06.1 + 524).

Results

The complete list of 15 questions, categorized into the three difficulty levels, is presented in Table 1. A total of 20 clinical experts and 13 patient representatives from various countries contributed to the evaluation. The full list of reviewers is provided in the Acknowledgments section.

Reliability

Reliability was assessed for all 15 questions submitted to the three AI tools. Out of 45 different outcomes evaluated (15 per software), 37 were deemed reliable, while 8 were regarded as unreliable. ChatGPT produced 14 reliable responses, Bard 11, and Copilot 12.

Accuracy

Accuracy was evaluated for all 15 questions administered to the three analysed software tools. Box and whisker plots depicting accuracy evaluations for all questions are displayed in Fig. 1. Median accuracy scores ranged between 6.5 (Question 3) and 9.0 (Question 5) for ChatGPT, between 5.5 (Question 3) and 7.0 (Questions 4, 5, 7, 9, 10, 14 and 15) for Bard and between 4.0 (Question 6) and 7.0 (Questions 1, 9, 10, 12, 13 and 15) for Copilot. Globally, accuracy scores of ChatGPT (median 8.0, IQR [7.0–9.0]) were higher than those of Bard (median 6.0, IQR [5.0–8.0]) and Copilot (median 6.0, IQR [5.0–7.3]) ($P < 0.0001$). No differences were detected when accuracy evaluations were compared according to the difficulty tiers of questions (ChatGPT: easy 8.0 [7.0–9.0] *vs.* medium 8.0 [7.0–9.0] *vs.* difficult 8.0 [6.0–9.0], $P = 0.5642$; Bard: easy 6.0 [5.0–8.0] *vs.* medium 7.0 [5.0–8.0] *vs.* difficult 6.0 [5.75–8.0], $P = 0.6551$; Copilot: easy 6.0 [5.0–8.0] vs. medium 6.0 [4.0–7.0] vs. difficult 6.50 [4.75–8.0], $P = 0.6814$). In all three difficulty tiers, ChatGPT's answers were more accurate than those of Bard and Copilot ones (easy: 8.0 [7.0–9.0] vs. 6.0 [5.0–8.0] vs. 6.0 [5.0–8.0], $P < 0.0001$; Medium: 8.0 [7.0–9.0] vs. 7.0 [5.0–8.0] vs. 6.0 [4.0–7.0], $P < 0.0001$; difficult: 8.0 [6.0–9.0] vs. 6.0 [5.75–8.0] vs. 6.50 [4.75–8.0], $P < 0.0001$).

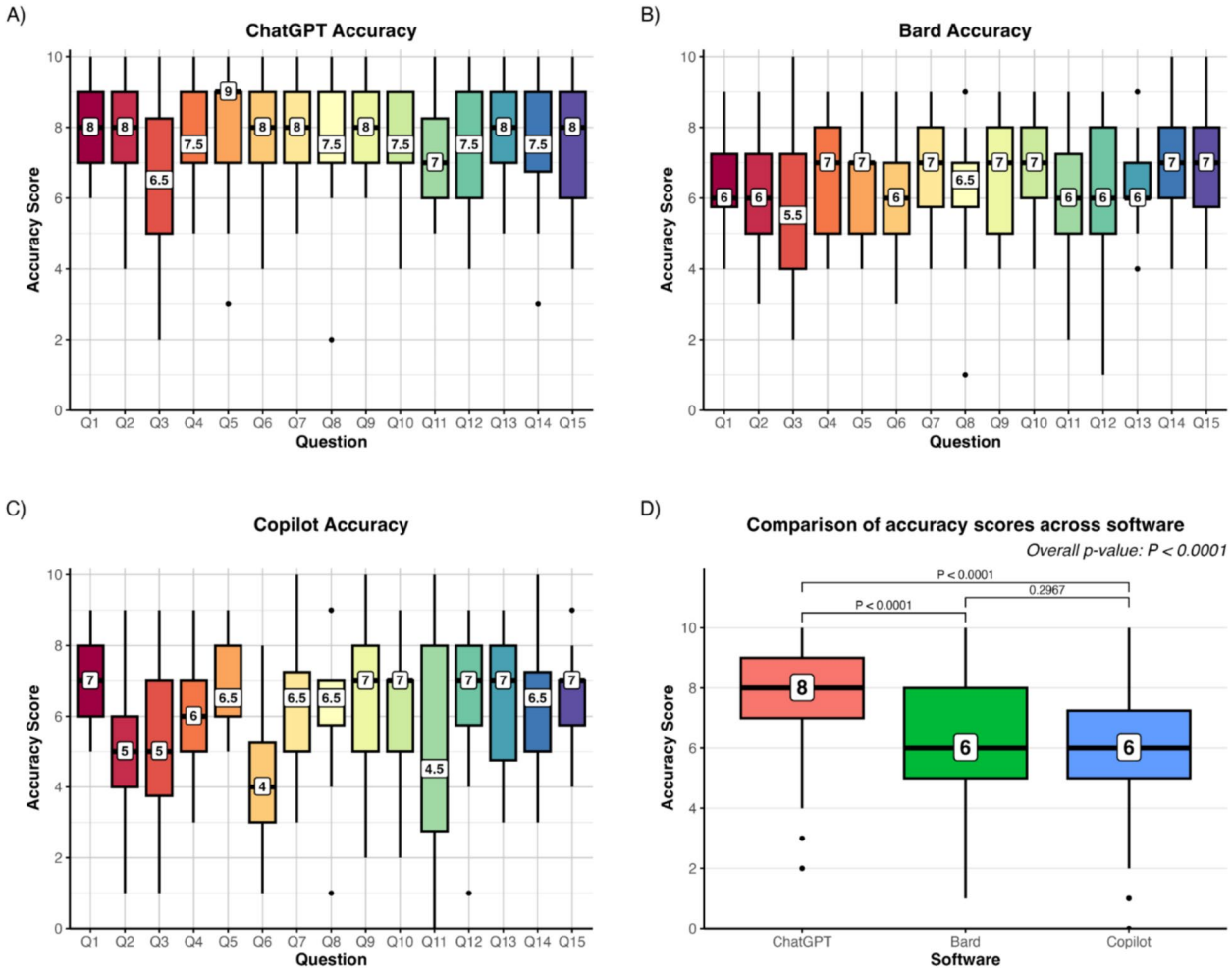

Fig. 1 Box and whiskers plots depicting accuracy evaluations for all questions. Displayed values are the medians of each distribution. **A** Accuracy evaluation for answers provided by ChatGPT; **B** Accuracy evaluation for answers provided by Bard; **C** Accuracy evaluation for answers provided by Copilot; **D** Comparison of accuracy evaluations between the three software

Comprehensiveness

Comprehensiveness was evaluated for all 15 questions administered to the three analysed software tools. Box and whisker plots showing comprehensiveness evaluations for all questions are displayed in Fig. 2. Median comprehensiveness scores ranged between 7.0 (Question 3) and 8.0 (Questions 1, 2, 4, 5, 6, 7, 9, 13 and 14) for ChatGPT, between 6.0 (Questions 1, 2, 6 and 7) and 7.0 (Questions 9, 13, 14 and 15) for Bard, and between 3.0 (Question 6) and 7.5 (Question 1) for Copilot. Comprehensiveness scores for answers provided by ChatGPT (8.0 [6.8–9.0]) were overall higher than the ones provided by both Bard (6.0 [5.0–7.0]) and Copilot (6.0 [5.0–8.0]) ($P < 0.0001$). No differences were detected when comprehensiveness evaluations were compared according to the difficulty tiers of questions (ChatGPT: easy 8.0 [6.8–9.0] *vs.* medium 8.0 [7.0–9.0] *vs.* difficult 8.0 [6.0–9.0], $P = 0.9791$; Bard: easy 6.0 [5.8–7.0] *vs.* medium 6.0 [5.0–7.25] *vs.* difficult 7.0 [5.8–8.0], $P = 0.6621$; Copilot: easy 6.0 [5.0–7.3] *vs.* medium 6.0 [5.0–7.3] *vs.* difficult 6.0 [5.0–8.0], $P = 0.4748$). In all three difficult tiers, ChatGPT answers were more comprehensive than those provided by Bard and Copilot (easy: 8.0 [6.8–9.0] vs. 6.0 [5.8–7.0] vs. 6.0 [5.0–7.3], $P < 0.0001$; medium: 8.0 [7.0–9.0] *vs.* 6.0 [5.0–7.3] *vs.* 6.0 [5.0–7.3], $P < 0.0001$; difficult: 8.0 [6.0–9.0] *vs.* 7.0 [5.8–8.0] *vs.* 6.0 [5.0–8.0], $P < 0.0001$).

Understandability

Understandability was evaluated for all 15 questions administered to the three analysed software tools. Box and whiskers plots displaying understandability evaluations for all provided questions are depicted in Fig. 3. Median understandability scores ranged between 8.0 (Questions 2, 3, 4, 10, 14 and 15) and 9.0 (all remaining questions) for ChatGPT, between 7.0 (Questions 12 and 14) and 10.0 (Questions 8 and 13) for Bard, and between 8.0 (Questions 2, 4, 5, 6, 7, 13 and 14) and 10.0 (Questions 1 and 11) for Copilot. Understandability scores

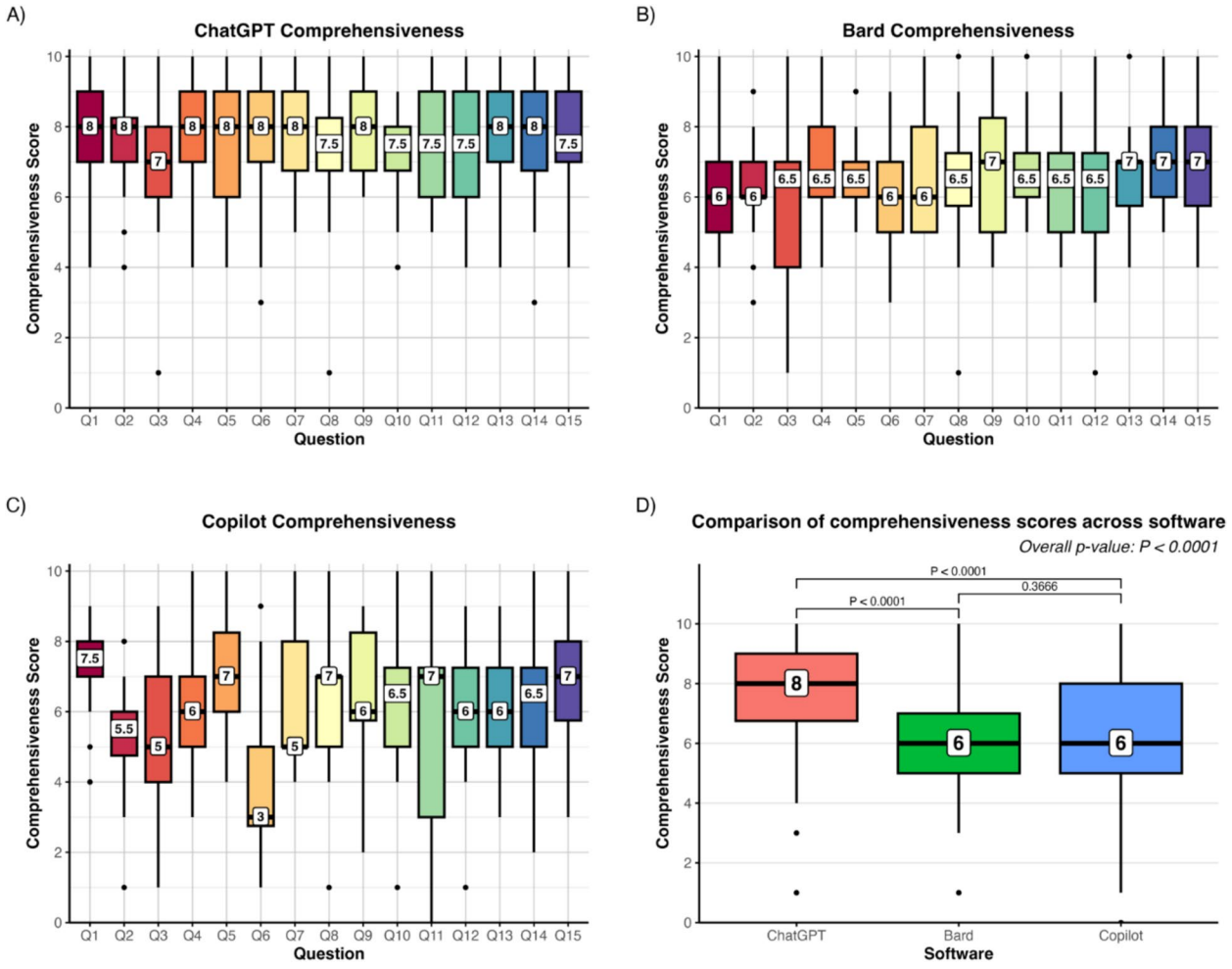

Fig. 2 Box and whiskers plots showing comprehensiveness evaluations for all questions. Displayed values are the medians of each distribution. **A** Comprehensiveness evaluations for answers provided by ChatGPT; **B** Comprehensiveness evaluations for answers provided by Bard; **C** Comprehensiveness evaluations for answers provided by Copilot; **D** Comparison of comprehensiveness evaluations between the three software

for answers provided by all three software were similar (ChatGPT: 8.0 [8.0–10.0]; Bard: 9.0 [8.0–10.0]; Copilot: 9.0 [8.0–10.0], overall $P = 0.5283$). No differences were detected when understandability evaluations were compared according to the difficulty tiers of questions, except for Bard (ChatGPT: easy 8.0 [7.0–10.0] *vs.* medium 9.0 [8.0–10.0] *vs.* difficult 9.0 [7.0–10.0], $P = 0.4351$; Bard: easy 9.0 [8.0–10.0] *vs.* medium 9.0 [8.0–10.0] *vs.* difficult 8.0 [7.0–10.0], $P = 0.01$; Copilot: easy 9.0 [8.0–10.0] *vs.* medium 9.0 [8.0–9.0] *vs.* difficult 9.0 [8.0–10.0], $P = 0.7114$). In all three difficulty tiers, understandability was similar across all of the software responses (Easy: ChatGPT 8.0 [7.0–10.0] *vs.* Bard 9.0 [8.0–10.0] *vs.* Copilot 9.0 [8.0–10.0], $P = 0.3619$; Medium: ChatGPT 9.0 [8.0–10.0] vs. Bard 9.0 [8.0–10.0] vs. Copilot 9.0 [8.0–9.0], $P = 0.0376$; Difficult: ChatGPT 9.0 [7.0–10.0] vs. Bard 8.0 [7.0–10.0] vs. Copilot 9.0 [8.0–10.0], $P = 0.4126$).

Discussion

Our findings demonstrate that ChatGPT, Bard, and Copilot performed overall adequately in responding to COPD-related questions formulated by patients. While none of the tools can be considered fully reliable, ChatGPT showed greater reliability than both Bard and Copilot. Moreover, ChatGPT consistently outperformed both Bard and Copilot in terms of accuracy and comprehensiveness, with no significant differences observed in terms of understandability. Notably, the performance of all three AI tools remained consistent across varying levels of question difficulty, suggesting independence from question complexity.

Although the analysis was limited to the free-access versions of the AI platforms, our results suggest that ChatGPT may currently be the most effective tool for generating clinically relevant, consistent, and detailed content in response to COPD-related queries. The lack of differences in understandability scores is particularly

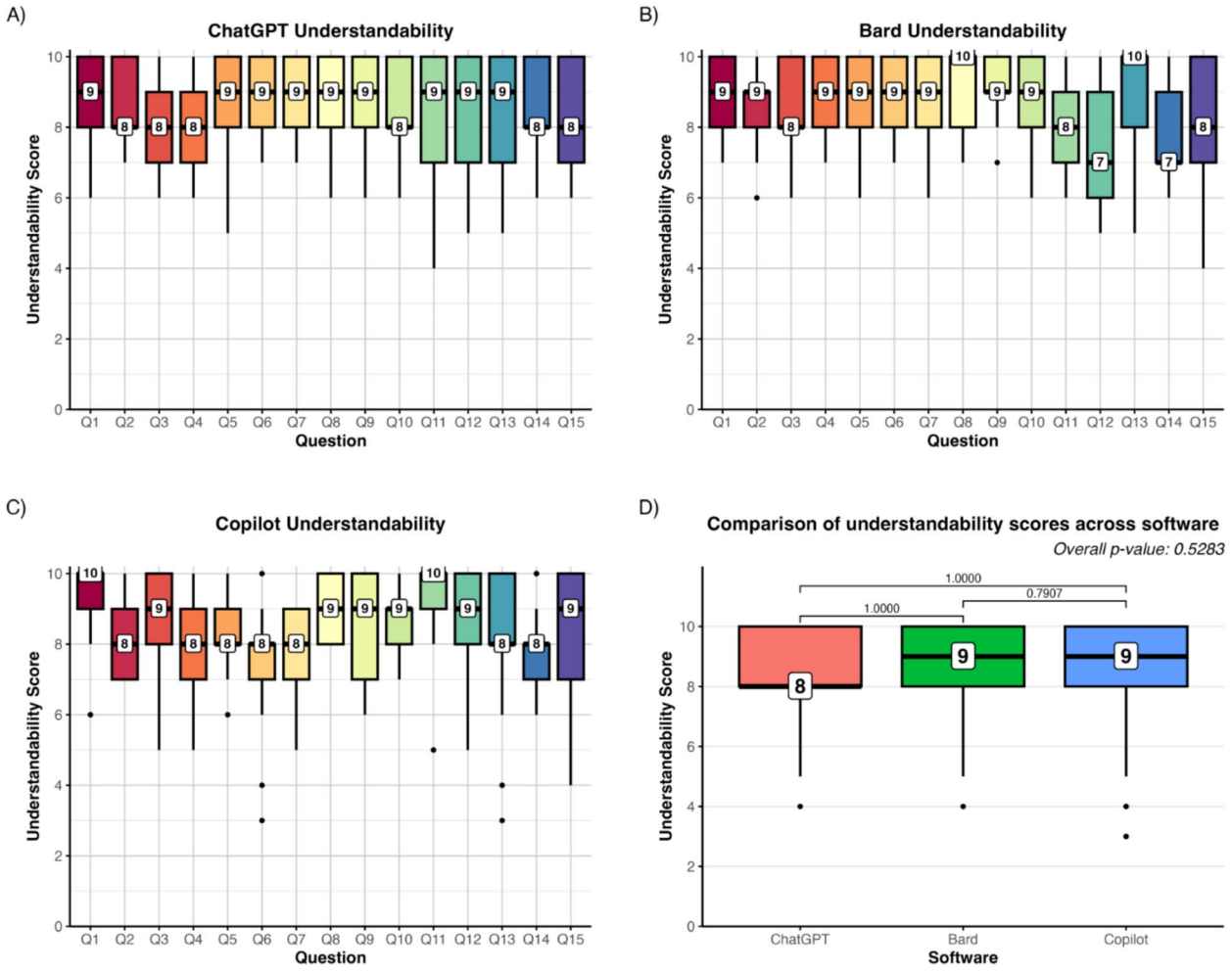

Fig. 3 Box and whiskers plots displaying understandability evaluations for all provided questions. Displayed values are the medians of each distribution. **A** Understandability evaluations for answers provided by ChatGPT; **B** Understandability evaluations for answers provided by Bard; **C** Understandability evaluations for answers provided by Copilot; **D** Comparison of understandability evaluations between the three software

notable from a patient-centered perspective, indicating that all three tools can produce content that is accessible and comprehensible to COPD patients. However, some concerns must be raised regarding reliability: although high (14 out of 15 for ChatGPT), a less-than-perfect reliability rate may be inadequate for patient-directed health information, as even a single inaccurate response—especially when delivered with the same confidence as a correct one—can pose significant risks.

To date, very few studies have investigated the role of AI in improving disease awareness, either as a direct-to-patient tool or as a support for healthcare professionals in educational settings. Jabeen and Saji found that educational materials generated by ChatGPT and DeepSeek across several chronic diseases, including COPD, did not meet recommended standards for readability, originality, or actionability, underscoring the need for human oversight [17]. Similarly, Yin et al. reported that although ChatGPT's responses to patient-centered COPD questions were highly reproducible and free of inaccuracies, they were frequently only partially accurate due to omissions and oversimplifications, supporting its use as a supplementary rather than standalone tool; this study did not include a direct comparison with other chatbots [18]. Another study comparing ChatGPT, Gemini, and DeepSeek showed that all three chatbots produced evidence-based COPD information of similar overall quality in response to non–patient-generated questions, with DeepSeek demonstrating better readability for several items and Gemini providing more academically complex content [19]. Our study expands this emerging literature by evaluating these aspects using real-world, patient-generated questions assessed by blinded COPD experts.

To our knowledge, only one other study has assessed ChatGPT 3.5 and Bing in the context of COPD clinical management [15]. However, that analysis had several limitations. First, the absence of patient involvement in that study makes it difficult to frame it as an educational

or disease awareness initiative, and the relevance of questions selected by researchers may not reflect actual patient concerns. Second, the formulations of questions by clinicians may have influenced the results, as clinical syntaxis may differ from patient language. Third, the inclusion of only UK participants limits the generalizability of the findings, as national or local healthcare issues may influence both questions content and interpretation. Fourth, the assessment was limited to only two AI chatbots, with no stratification for question difficulties. Finally, the use of a 5-point evaluation scale, restricting the granularity of the assessments, and the choice of statistical tests requiring normal distributions have potentially impacted outcomes. In contrast, our results align closely with those of a recent study conducted on asthma employing the same methodology [14]. In that study, ChatGPT also emerged as the most reliable, accurate and comprehensive tool, with Bard rated highest in terms of understandability. Furthermore, as in our analysis, no effect of question difficulty on AI chatbots performances was detected. This concordance strengthens the validity of our findings and highlights the growing relevance of AI tools in respiratory care.

From a clinical perspective, our results suggest that generative AI models, particularly ChatGPT, may serve as useful adjuncts in providing reliable health information to patients with COPD. While these tools should not be considered substitutes for medical advice, they may help clarify key concepts and address patient concerns, thereby indirectly supporting treatment adherence through improved understanding and engagement. This is especially relevant in contemporary healthcare systems, where time and personnel constraints often limit in-depth communication between clinicians and patients. Improving disease awareness in COPD has the potential to empower patients, promote earlier diagnosis, and support lifestyle changes and self-management plans, all of which may contribute to improved health outcomes. By providing trustworthy, patient-accessible information, generative AI tools can help individuals recognize symptoms earlier, understand their disease better, and take a more active role in disease management. In a broader sense, this study serves as a foundational model for evaluating the performance of LLM- based tools and services in healthcare. Our results address a key gap in the current literature and highlight the need for ongoing research. Future studies should include larger and more diverse question sets, involve broader patient populations, assess premium versions of LLM-based tools and services, and employ longitudinal designs to monitor evolving model performance. Additionally, investigations should explore the real-world impact of AI-generated information on patient behavior, clinical decision-making, and health outcomes.

This work has several strengths. First, it was conducted as a collaborative international initiative involving two prominent European institutions (ELF and ERS), providing methodological rigor and broad relevance. Second, the study actively involved people with COPD in developing the questions, ensuring that the evaluated content was authentic, reflective of real-world concerns, and aligned with actual patient information needs. The questions were not edited or rephrased by the investigators, preserving both the linguistic authenticity and the patient voice. Third, we engaged a diverse group of international COPD experts to assess accuracy and comprehensiveness, and experienced ELF patient representatives to evaluate understandability. This dual-review process enabled a balanced, multidisciplinary, and user-centered assessment of the responses. Fourth, the use of multiple AI tools allowed for an expansive comparative analysis, capturing variability in performance that may not be detectable in single-model studies. The stratification of questions by difficulty level further enhanced the robustness of our design, enabling assessment of model performance across a spectrum of complexity. Additional methodological strengths included blinded evaluations and the inclusion of only fully completed assessments, which minimized potential biases and strengthened the generalizability of our findings.

Nonetheless, several limitations must be acknowledged. The cross-sectional design did not allow for assessment of temporal performance variation, especially considering that AI software may have increased their performance with updates. Furthermore, we employed the free version of ChatGPT, which may differ in capabilities from its paid premium counterpart, which allows user to exploit the most recent versions available. However, this choice reflects real-world usage patterns, as most patients are unlikely to subscribe to paid versions when free options are readily available and sufficiently responsive. Another limitation was the relatively small number of questions analyzed, which will not capture the full range of patient information needs. Furthermore, question generation and understandability evaluation was performed only by individuals with a confirmed diagnosis of COPD, and it remains uncertain whether our findings apply to people with compatible symptoms but no diagnosis. Also, no comparison between AI-generated answers and the ones that experts would have provided in similar circumstances has been performed. Still, because the questions were collected from members of the COPD PAG of the ELF, participants may have had higher health literacy than the broader COPD population, potentially introducing a selection bias in both the formulation of questions and the understandability assessment. Lastly, although the evaluation process was

blinded, the classification of question difficulty was based on investigator judgment and thus remains subjective.

In summary, our findings demonstrate that AI-based tools, and particularly ChatGPT, can provide accurate, comprehensive, and understandable answers to patient-generated questions about COPD. While these tools hold promise as user-friendly and scalable solutions to support patient education, their artificial and non-personalized nature, as well as their lack of empathetic communication, highlight that they should be viewed as adjuncts rather than substitutes for human interaction within patient-centred care frameworks.

Abbreviations

AI	Generative artificial intelligence
COPD	Chronic obstructive pulmonary disease
CRC	Clinical research collaboration
ELF	european lung foundation
ERS	European respiratory society
IP	Internet protocol
IQR	Interquartile range
LLM	Large language model
PAG	Patient advisory group
RNG	Random number generated

Acknowledgements
We acknowledge the support of ELF administration, the European Lung Foundation COPD patient advisory group and their caregivers, and all the experts recruited through the European Respiratory Society and the ERS Clinical Research Collaboration CONNECT who dedicated time to this initiative. Furthermore, we acknowledge Professor Alessio Aghemo and Doctor Nicola Pugliese (Department of Biomedical Sciences, Humanitas University, Milan, Pieve Emanuele, Italy; Division of Internal Medicine and Hepatology, Department of Gastroenterology, IRCCS Humanitas Research Hospital, Milan, Rozzano, Italy.).
The following patients contributed to the evaluations: Phil Collis, Gilberto Calderoni, Uwe Schmitt, Ariane Bruinen, Claire, Lara Consani, Matt Cullen, Helen Parks, Glo Wils, Tessa Jelen, Rikki Muller, Carol Hearson, Linda Clephane
We acknowledge the support of ELF administration, the European Lung Foundation COPD patient advisory group and their caregivers, and all the experts recruited through the European Respiratory Society and the ERS Clinical Research Collaboration CONNECT who dedicated time to this initiative. Furthermore, we acknowledge Professor Alessio Aghemo and Doctor Nicola Pugliese (Department of Biomedical Sciences, Humanitas University, Milan, Pieve Emanuele, Italy; Division of Internal Medicine and Hepatology, Department of Gastroenterology, IRCCS Humanitas Research Hospital, Milan, Rozzano, Italy.). This work was partly funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022)—project n. PNC0000003—AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

The AIR-COPD Task Force consortium was formed by
Antonio Spanevello[1,2], Marco Vanetti[1,2], David MG Halpin[3], Khanh Le Quoc Tran[4], Maria Elia Gomez-Merino[5], Deepak Muthreja[6], Boudewijn J.H. Dierick[7], Elene Khurtsidze[8], Vishakha Kalpesh Kapadia[9], Amalia Panagiotou[10], Miguel Gallego[11,12], Stavros Tryfon[13], Efthymia Papadopoulou[13], Carlos Figueiredo[14], Shailesh Balasaheb Kolekar[15], Pradeesh Sivapalan[16], Pedro J. Marcos[17].

Authors' contributions
Study conception and design: MN and SA. Data collection and recruitment: MN, AB, HP, JB, PP and SA. Data analysis: MN and SA. Writing of the manuscript: MN, GEB and SA. Revising the manuscript for important content: MN, GEB, And Ali, Ale Ang, AKS, Ant Anz, PMC, FA, AS, AB, HP, JB, PP, SA. Figure drafting: MN.

Funding
This work was partly funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06–06-2022)—project n. PNC0000003—AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

Data availability
Reviewers' evaluations have been exported into an Excel spreadsheet and are available for consultation, if needed.

Declarations
The AIR-COPD study was conducted in agreement with the current version of Declaration of Helsinki and the applicable regulations.

Ethics approval and consent to participate
Not applicable. No interventions have been performed on patients for this research.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, Pieve Emanuele, 20072 Milan, Italy
[2]IRCCS Humanitas Research Hospital, Respiratory Unit, Via Manzoni 56, 20089 Rozzano, Milan, Italy
[3]Dipartimento Di Elettronica, Informazione E Bioingegneria, Politecnico Di Milano, Piazza Leonardo Da Vinci 32, 20133 Milan, Italy
[4]Sleep and Ventilation Unit, Royal Brompton and Harefield Hospital (Guys and St Thomas' NHS Foundation Trust), London, UK
[5]National Heart and Lung Institute, Imperial College London, London, UK
[6]South Texas Veterans Health Care System, University of Texas, San Antonio, TX, USA
[7]Institute of Life Course and Medical Sciences, University of Liverpool, Liverpool, UK
[8]Karolinska Severe Asthma Centre, Department of Respiratory Medicine and Allergy, Karolinska University Hospital, Stockholm, Sweden
[9]Division of Lung and Airway Research, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[10]Lung Laboratory, Centre for Molecular Medicine, Karolinska University Hospital, Stockholm, Sweden
[11]Usher Institute, The University of Edinburgh, Edinburgh, UK
[12]European Lung Foundation, Sheffield, UK

Received: 14 September 2025 / Accepted: 20 November 2025
Published online: 16 December 2025

References

1. Perret J, Yip SWS, Idrose NS, et al. Undiagnosed and 'overdiagnosed' COPD using postbronchodilator spirometry in primary healthcare settings: a systematic review and meta-analysis. BMJ Open Respir Res. 2023;10:e001478.
2. Yousuf A, McAuley H, Elneima O, et al. The different phenotypes of COPD. Br Med Bull. 2021;137:82–97.
3. Cardoso J, Ferreira AJ, Guimarães M, et al. Treatable traits in COPD – a proposed approach. Int J Chron Obstruct Pulmon Dis. 2021;16:3167–82.
4. Esam Mahmood SA, Alqahtani AT, Alghamdi BAA, et al. Awareness of COPD and its risk factors among the adult population of the Aseer Region Saudi Arabia. Int J Chron Obstruct Pulmon Dis. 2023;18:23–35.
5. Schrijver J, Lenferink A, Brusse-Keizer M, et al. Self-management interventions for people with chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2022;2023.

6. Baiardini I, Rogliani P, Santus P, et al. Disease awareness in patients with COPD: measurement and extent. Int J Chron Obstruct Pulmon Dis. 2018;14:1–11.
7. Baiardini I, Contoli M, Corsico AG, et al. Exploring the relationship between disease awareness and outcomes in patients with chronic obstructive pulmonary disease. Respiration. 2021;100:291–7.
8. Ghorpade D, Salvi S. Awareness of COPD in low-and middle-income countries and implications for treatment. Expert Rev Respir Med. 2024;18:721–33.
9. Kurniawan MH, Handiyani H, Nuraini T, et al. A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. Ann Med. 2024. https://doi.org/10.1080/07853890.2024.2302980.
10. Pugliese N, Wai-Sun Wong V, Schattenberg JM, et al. Accuracy, reliability, and comprehensibility of ChatGPT-generated medical responses for patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol. 2024;22:886-889.e5.
11. Goodman RS, Patrinely JR, Stone CA, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6:e2336483.
12. Kassab J, Hadi El Hajjar A, Wardrop RM, et al. Accuracy of online artificial intelligence models in primary care settings. Am J Prev Med. 2024;66:1054–9.
13. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183:589.
14. Nigro M, Aliverti A, Angelucci A, et al. Artificial Intelligence-generated answers to patients' questions on asthma: the AIR-Asthma study. J Allergy Clin Immunol Pract 2025.
15. Imtiaz A, King J, Holmes S, et al. ChatGPT *versus* Bing: a clinician assessment of the accuracy of AI platforms when responding to COPD questions. Eur Respir J. 2024;63:2400163.
16. European Respiratory Society. CONNECT – Available at: - Available at: https://www.ersnet.org/science-and-research/clinical-research-collaboration-application-programme/connect-moving-multiple-digital-innovations-towards-connected-respiratory-care-addressing-the-over-arching-challenges-of-whole-systems-implementation/.
17. Jabeen J, Saji JG. Evaluating AI-generated patient education guides: a comparative study of ChatGPT and Deepseek. Cureus 2025
18. Yin Y, Riaz Z, Amoro Sanchez R, et al. Evaluating ChatGPT as a standalone tool for patient education: a review of frequently asked questions by patients with chronic obstructive pulmonary disease. Cureus. 2025. https://doi.org/10.7759/cureus.92519.
19. Merç P, Pirinççi CŞ, Cihan E. Evaluation of AI chatbots for patient education and information on chronic obstructive pulmonary disease. Heart Lung. 2026;75:21–5.

Publisher's Note