

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:<https://orca.cardiff.ac.uk/id/eprint/172097/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Imtiaz, Arouba, King, Joanne, Holmes, Steve, Gupta, Ayushman, Bafadhel, Mona, Melcher, Marc L., Hurst, John R., Farewell, Daniel, Bolton, Charlotte E. and Duckers, Jamie 2024. ChatGPTversusBing: a clinician assessment of the accuracy of AI platforms when responding to COPD questions. *European Respiratory Journal* 63 (6), 2400163. 10.1183/13993003.00163-2024

Publishers page: <http://dx.doi.org/10.1183/13993003.00163-2024>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



ChatGPT vs. Bing: Clinicians Assessment of the Accuracy of AI Platforms in Responding to COPD Questions

Arouba Imtiaz¹, Joanne King², Steve Holmes³, Ayushman Gupta⁴, Mona Bafadhel⁵, Marc L. Melcher⁶, John R Hurst⁷, Daniel Farewell⁸, Charlotte E Bolton⁹, Jamie Duckers¹⁰

¹ University Hospital of Wales, Cardiff. ORCID: 0009-0005-3281-5903

² Respiratory Department, Frimley Health NHS Foundation Trust, UK

³ General Practitioner, The Park Medical Practice, Shepton Mallet, Somerset, UK

⁴ Respiratory Department, Nottingham University Hospital NHS Trust, Nottingham, UK. ORCID 0000-0002-2345-545X

⁵ King's Centre for Lung Health, School of Immunology and Microbial Sciences. King's College London, UK. ORCID: 0000-0002-9993-2478

⁶ Department of Surgery, Stanford University, USA

⁷ UCL Respiratory, University College London, London, UK. ORCID: 0000-0002-7246-6040

⁸ Professor of Statistics, Cardiff University, Wales UK

⁹ Centre for Respiratory Research, NIHR Nottingham Biomedical Research Centre, School of Medicine, University of Nottingham, Nottingham, UK, ORCID 0000-0002-9578-2249

¹⁰ All Wales Adult Cystic Fibrosis Centre, Wales, UK

Word Count: 1,394 words (excluding title page, table and references)

Chronic obstructive pulmonary disease (COPD) is a global respiratory challenge ranking third in morbidity and mortality, primarily caused by exposure to particulate matter, notably from tobacco smoke (1, 2). The resulting inflammation and lung damage can lead to symptoms of symptoms such as shortness of breath, wheezing, cough and sputum production (1).

With more emphasis on patient-centred care and greater accessibility of information available through online search engines, many people, including those living with COPD are turning to online platforms to seek further guidance. However, there is a concern that the quality of information provided may not be regulated and may lead to sub-optimal outcomes (2). In response, it transpires that Google has adjusted its algorithms over time to reduce the visibility of medical websites that do not meet stringent ranking criteria, aiming to mitigate the spread of misinformation (3).

The rise of artificial intelligence (AI), large language models, such as ChatGPT, (developed by OpenAI and released in November 2022), and Microsoft Bing Chat (released in May 2023) is already impacting the field of medicine (4, 5). ChatGPT has been reported to surpass Google Search in terms of medical knowledge (6) and it can be challenging for patients to distinguish AI advice from clinical providers' advice (7). However, the lack of formal assessment for information from AI platforms demonstrates the need to evaluate the accuracy of content disseminated through these channels.

This research aimed to assess the appropriateness of AI responses to COPD-related medical questions, evaluating their suitability, from the perspective of clinicians involved in COPD care.

The researchers generated 21 questions to reflect clinicians' views on common patient inquiries during clinic visits, covering topics such as the nature of the condition, treatment options, investigations, potential complications, and general queries. See Table 1b for the specific questions.

Each question was inputted into two AI platforms: ChatGPT version 3.5 and Microsoft Bing Chat, and the responses were recorded (4, 5). The hyperlinks were removed from the Microsoft Bing Chat's responses, relying solely on the short response. The study, conducted in a blinded manner, involved respiratory specialists from the United Kingdom (UK); evaluating each response on a scale of 1 to 5, considering accuracy, completeness, clarity, and safety. Parameters were defined as follows:

- **Accuracy:** The correctness of the response based on clinicians' opinions.
- **Completeness:** Evaluates whether the response provides all the necessary and important details.
- **Clarity:** Assesses the extent to which the response was easily understandable from the perspective of patients, avoiding medical jargon and ensuring comprehension among the target audience.
- **Safety:** Determining whether the advice given in the response is safe for the patient.

Microsoft Excel was used to summarise the responses, calculating mean and corresponding standard deviation across all the questions within a particular parameter, and for a particular platform. For each question, we computed the mean scores for both platforms, and visually

confirmed that these means had a roughly normal distribution. We therefore carried out a paired t-test on the resulting averages, and reported its p-value.

This study presents a comprehensive assessment of two commonly used AI platforms, ChatGPT version 3.5 and Microsoft Bing, in responding to COPD-related queries. The findings indicate that ChatGPT generally outperformed Microsoft Bing across key parameters—accuracy, completeness, clarity, and safety, most strikingly in completeness (see Table 1a). Clarity had the smallest difference, likely influenced by Microsoft Bing's shorter responses. Also, the Flesch-Kincaid Grade level was utilized to assess readability and clarity (8). ChatGPT's level was 13, while Microsoft Bing's was 10.8, indicating ChatGPT's higher complexity (8). This aligns with our study's conclusions. The lower standard deviation with ChatGPT indicates that across all 21 questions there was less variation in the scoring of responses for ChatGPT compared to Microsoft Bing Chat.

The comments on the responses to the questions brought attention to several notable themes. Microsoft Bing's responses were observed to be more concise compared to ChatGPT. The advice appeared to often be United States (US)-centric, relying on US guidelines and incorporating American drug names. Information was outdated in places, as there was no mention of vaping when advising on tobacco dependency strategies, nor was the importance of coronavirus (COVID-19) vaccines included. Another recurring issue in the responses was incomplete information, particularly concerning medications such as inhaled corticosteroids, long-acting beta-agonist and long-acting muscarinic-agonist used together as 'triple therapy'. The description of inhaler use was confined to metered-dose inhalers (MDIs), neglecting other types of inhalers. Additionally, there was misinformation about the use of oxygen therapy for acute breathlessness instead of inhalers.

Some safety concerns noted in the study include the absence of warnings regarding smoking while using long-term oxygen therapy, failure to address the risks associated with using oxygen in type 2 respiratory failure, and specific inhaler technique advice provided solely for MDIs, rendering it inadequate for other types of inhalers.

According to a meta-analysis, ChatGPT generally exhibits an accuracy of 56% for medical questions (9). Specifically, in the field of internal medicine, the meta-analysis indicated a higher accuracy of 63%, compared to 49% in surgery (9). A more recent publication, utilizing ChatGPT for cardiovascular disease advice, judged that responses were found to be 84% appropriate, as evaluated by cardiology clinicians (10). Interestingly, our study reveals a similar accuracy rate for COPD, with ChatGPT (lower for Bing). This suggests that ChatGPT may offer heightened accuracy in addressing specific medical conditions. Alternatively, these consistent outcomes may indicate an improvement in ChatGPT's accuracy in handling medical questions over time.

In our study, it was noted that there was an absence of certain key topics of information in the responses on important topics. This underscores one of the challenges with ChatGPT: they are constructed using past online information and do not adapt based on current information and updates. Consequently, with evolving guidelines and new treatments for medical conditions, AI platforms that do not update their source material may disseminate inaccurate information. A potential solution to this issue could involve creating a chatbot using the ChatGPT application programming interface (API) and incorporating medical guidelines specific to conditions. This approach would ensure that the information is up-to-date and accurate while responding to queries in a manner similar to ChatGPT.

Despite having more up-to-date information, Microsoft Bing also missed details about these topics, possibly due to its shorter responses. While these AI platforms offer reasonable clinical information, it's crucial to consider the user demographic. COPD patients, often being older and potentially less computer literate, may face challenges in utilizing these platforms effectively. However, it is still important to determine the level of accuracy of these AI platforms to consider their usage for patients.

In the realm of respiratory medicine, AI has been demonstrated to outperform respiratory physicians in interpreting pulmonary function tests (11). Additionally, AI models have shown the capability to predict the risk of developing asthma and COPD (12). This study contributes to the growing body of research in respiratory medicine and proposes the use of AI as a clinical decision support system to augment the work of doctors.

The study has certain limitations. Firstly, the questions were framed clinically, potentially influencing AI responses as they may not align with how a person living with COPD experiencing symptoms typically asks questions. This study was an initial exploration, and the next step would be to co-construct questions with a lay audience and seek their view alongside clinicians. Additionally, varying the presentation of questions could aid in drawing more comprehensive conclusions. However, reassuringly, views of clinicians that the advice is not harmful were deemed important. Furthermore, the judging was based on UK healthcare services. Lastly, the study used a small sample of questions specific to one condition, limiting the generalizability of the results to other medical conditions.

In conclusion, insights into the performance of AI platforms, specifically ChatGPT and Microsoft Bing, in responding to COPD-related medical queries are presented, with ChatGPT exhibiting superior performance in key parameters of accuracy, completeness, clarity, and safety. The substantial difference in completeness, with ChatGPT offering a more detailed response, underscores its potential to deliver accurate medical information. As AI continues to develop, platforms like ChatGPT will be valuable resources for delivering medical information, benefiting both patients and healthcare providers equally. Provided they remain up-to-date with guidance and tailor their responses to the specific healthcare services patients are exposed to or cared for by. Clinicians need to be mindful of these considerations when utilizing such AI platforms.

Table 1: Mean, Standard Deviation and t-Test of Scores for Each Parameter Across AI Platforms and Questions Asked in the Study

Table 1a: Mean, Standard Deviation and t-Test of Scores for Each Parameter Across AI Platforms

	Accuracy	Completeness	Clarity	Safety
ChatGPT (Out of 5)				
Mean (Standard Deviation)	4.34 (0.52)	4.45 (0.54)	4.49 (0.35)	4.41 (0.49)
Microsoft Bing Chat (Out of 5)				
Mean (Standard Deviation)	3.89 (0.61)	3.25 (0.72)	4.21 (0.53)	3.89 (0.63)
t-Test for Questions* Mean Difference	0.46	1.20	0.28	0.52

(Standard Deviation)	(0.51)	(0.68)	(0.52)	(0.54)
p-value	0.0006	<0.0001	0.025	0.0003

Table 1b: Questions Asked in the Study

***Questions:**

1. What is COPD?
2. Are there any environmental triggers that can worsen my COPD?
3. How do I quit smoking with COPD?
4. How do I properly use my inhaler?
5. What vaccinations do I need when diagnosed with COPD?
6. How can I prevent respiratory infection with COPD?
7. What exercises can I do with COPD?
8. Which medications should I take when I am short of breath with COPD?
9. What do I do if I am still short of breath after using my inhalers with COPD?
10. Can I use oxygen to help with my COPD?
11. Are there any alternative therapies other than my inhaler for COPD?
12. What are the side effects of my COPD medications?
13. How often should I see my doctor for check-ups for COPD?
14. What are the potential complications of COPD?
15. How can I manage my mucus production related to COPD?
16. How can I improve my lung function with COPD?
17. How can I manage my anxiety related to COPD?
18. What is the prognosis for someone with COPD?
19. How can I manage my COPD while traveling?
20. What is spirometry?
21. What is pulmonary rehabilitation?

References

1. Christenson SA, Smith BM, Bafadhel M, Putcha N. Chronic obstructive pulmonary disease. *The Lancet* [Internet]. 2022 May 6;399(10342). Available from: <https://www.sciencedirect.com/science/article/pii/S0140673622004706>
2. Fang Y, Shepherd TA, Smith HE. Examining the Trends in Online Health Information–Seeking Behavior About Chronic Obstructive Pulmonary Disease in Singapore: Analysis of Data From Google Trends and the Global Burden of Disease Study. *Journal of Medical Internet Research*. 2021 Oct 18;23(10):e19307.
3. Strzelecki A. Google Medical Update: Why Is the Search Engine Decreasing Visibility of Health and Medical Information Websites? *International Journal of Environmental Research and Public Health*. 2020 Feb 12;17(4):1160.
4. OpenAI. ChatGPT [Internet]. openai.com. 2023. Available from: <https://openai.com/chatgpt>
5. Microsoft. Your AI-powered Copilot for the Web | Microsoft Bing [Internet]. www.microsoft.com. Available from: <https://www.microsoft.com/en-us/bing?ep=0&es=31&form=MA13FV>
6. Ayoub N, Lee Y, Grimm DR, Vasu Divi. Head-to-Head Comparison of ChatGPT Versus Google Search for Medical Knowledge Acquisition. *Otolaryngology-Head and Neck Surgery*. 2023 Aug 2;
7. Nov O, Singh N, Mann D. Putting ChatGPT’s Medical Advice to the (Turing) Test (Preprint). *JMIR medical education*. 2023 Jul 10;9:e46939–9.

8. Flesch Kincaid Calculator [Internet]. 2024. Available from: <https://goodcalculators.com/flesch-kincaid-calculator/>
9. Wei Q, Yao Z, Ying C, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-Generated Medical Responses: A Systematic Review and Meta-Analysis. arXiv (Cornell University). 2023 Oct 12;
10. Ashish Sarraju, Bruemmer D, Van EH, Cho L, Rodriguez F, Laffin LJ. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA. 2023 Feb 3;329(10):842–2.
11. Topalovic M, Das N, Burgel P-R, Daenen M, Derom E, Haenebalcke C, et al. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. European Respiratory Journal. 2019 Feb 14;53(4):1801660. doi:10.1183/13993003.01660-2018
12. Hernandez R, Rodriguez R, Villalva O, Pimentel J, Miguel JL, Villasis MA. Exploratory study of a risk prediction artificial intelligence model to diagnose asthma and COPD in Mexico. Monitoring airway disease. 2023 Sept 9; doi:10.1183/13993003.congress-2023.pa3784