Chronic Obstructive Pulmonary Diseases:
Journal of the COPD Foundation®

Original Research

Accuracy, Usefulness, and Impact Variability of ChatGPT-4o for COPD Medication Management: A Modified Delphi Study

Paul M. Boylan, PharmD[1] Devin L. Lavender, PharmD[2] Rebecca H. Stone, PharmD[2] Russ Palmer, PhD[2]
Richard L. Lamb, PhD[2,3] Kiley Merritt, BS[2] Joshua Caballero, PharmD[2]

Abstract

Background: Chronic obstructive pulmonary disease (COPD) management is complex and rapidly evolving. ChatGPT is a large language model (LLM) shown to generate treatment plans for chronic conditions, yet its accuracy, usefulness, and consistency for COPD remain poorly characterized. This study evaluated the accuracy, usefulness, and impact variability of simultaneous ChatGPT-4o responses to COPD medication management questions.

Methods: Five COPD treatment questions were simultaneously entered into 3 separate computers using ChatGPT-4o during a single session, generating 15 total responses. Three residency-trained, board-certified clinical pharmacists rated each response across 3 domains—accuracy, usefulness, and impact variability—using a 3-point ordinal scale (0 to 2) via a 3-round modified Delphi process. Consensus was defined *a priori* as unanimous agreement among all 3 panelists.

Results: Of 45 response-domain ratings, consensus was achieved in 40 (88.9%). Accuracy ranged from poor (0) to good (2), usefulness from somewhat (1) to very useful (2), and impact variability from low (0) to high (2). For one question on stable COPD pharmacotherapy, all 3 simultaneous responses cited a retired clinical practice guideline, resulting in poor accuracy ratings. For 2 questions—treating exacerbations and managing a complex case—one response per question was rated higher in usefulness than the others.

Conclusions: Simultaneous ChatGPT-4o responses to identical COPD prompts differed materially in accuracy and potential clinical impact. LLM outputs should augment rather than replace clinician judgment and must be deployed with current guideline grounding and appropriate local oversight.

1. College of Pharmacy, University of Oklahoma, Oklahoma City, Oklahoma, United States
2. College of Pharmacy, University of Georgia, Athens, Georgia, United States
3. College of Veterinary Medicine, University of Georgia, Athens, Georgia, United States

Abbreviations:

AI=artificial intelligence; **CI**=confidence interval; **COPD**=chronic obstructive pulmonary disease; **ECOPD**=exacerbation of COPD; **GOLD**=Global initiative for chronic Obstructive Lung Disease; **LLM**=large language model

Funding Support:

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sections.

Citation:

Boylan PM, Lavender DL, Stone RH, et al. Accuracy, usefulness, and impact variability of ChatGPT-4o for COPD medication management: a modified Delphi study. ***Chronic Obstr Pulm Dis.*** 2026;13(5):386-394. doi: https://doi.org/10.15326/jcopdf.2026.0794

Publication Dates:

Date of Acceptance: July 29, 2026
Published Online Date: August 4, 2026

Address correspondence to:

Paul Boylan, PharmD
1110 N Stonewall Ave CPB 239
Oklahoma City, OK, 73117
Phone: (405) 271-6878 x47255
Email: paul-boylan@ou.edu

Keywords:

COPD; ChatGPT; artificial intelligence; large language models; clinical decision-making

This article has an online supplement.

Note: A version of these results was previously presented as a poster at the American College of Clinical Pharmacy Annual Meeting in Minneapolis, Minnesota in October 2025.

For personal use only. Permission required for all other uses.

Introduction

Chronic obstructive pulmonary disease (COPD) affects an estimated 392 million people, or 6% of all global deaths, with prevalence highest among adults greater than 40 years of age, particularly in low and middle income countries where exposure to tobacco smoke, biomass fuel, and air pollution is common.[1-5] Large population-based analyses estimate global prevalence among individuals aged 30–79 years at 10.3%, with substantial regional variation driven by differences in smoking rates, occupational exposures, and socioeconomic status.[6,7] COPD prevalence and mortality are rising among women, particularly where smoking and household air pollution are prevalent.[8] Projections suggest that, despite advances in prevention and treatment, COPD prevalence and mortality will continue to rise through 2050, particularly in aging populations and regions with persistent exposure to modifiable risk factors.[8]

In addition to rising prevalence, patients lack access to adequately trained health care professionals, particularly in rural and resource-limited settings.[9] As a result, patients and providers lacking training and experience in COPD may resort to alternative resources for disease management support, including artificial intelligence (AI)-based tools such as large language models (LLMs) to obtain information and guidance outside of traditional clinical care pathways. Ultimately, AI may aid clinicians to deliver patient care by collecting vast and disparate data to support accurate disease diagnosis, clinical reasoning, and treatment decisions.[10] AI could streamline clinicians' workload and enable health care providers to dedicate more time to patient care responsibilities.[11] Over the last decade, the U.S. Food and Drug Administration's public list of AI and machine learning-enabled medical devices has grown rapidly, with over 1300 authorizations across specialties as of the latest update.[10,12] Asian and European regulatory agencies have similarly convened focus groups, developed toolkits, and enhanced pharmacovigilance programs to facilitate patient safety, increase health care professional awareness, and minimize AI risks.[12]

In controlled comparisons, AI software has outperformed pulmonologists in pulmonary function test interpretation accuracy and diagnostic assessment, underscoring both the promise and need for careful guardrails in respiratory decision support.[13] ChatGPT (OpenAI; San Francisco, California) is an LLM AI that has demonstrated the potential to answer medical questions, including questions regarding COPD.[14-16] In a study by Imtiaz et al, ChatGPT responses demonstrated high degrees of accuracy, completeness, clarity, and safety when prompted to 21 COPD questions, including 6 items addressing COPD medications.[14] However, other studies have identified varying degrees of ChatGPT's accuracy responding to complex medication-related questions and noted low inter-response agreement when the same prompts are administered over time.[17,18] These aforementioned studies discuss single responses at one point in time or several responses over a period of time; however, limited data exist exploring differences between responses when the same question is simultaneously inputted across different devices using the same LLM AI (e.g., ChatGPT).[14-16] Given the prevalence and clinical burdens of COPD, further research on the accuracy, utility, and reliability of ChatGPT responding to COPD questions has been suggested.[10,19,20] These recent studies show mixed performance of LLMs on COPD tasks: in a clinician assessment of 21 COPD questions, ChatGPT provided detailed—but at times outdated—advice compared with Bing, while pharmacy-focused evaluations demonstrated limited accuracy and reproducibility on drug information queries, directly motivating ongoing needs to assess simultaneous LLM outputs.

The purpose of this study was to evaluate the accuracy, usefulness, and impact variability of ChatGPT-4o across 3 simultaneous responses across different computer devices and to describe reasons for differences in accuracy, usefulness, and impact variability performance.

Methods

This project used a 3-round modified Delphi approach with the purpose of forming consensus on ChatGPT's accuracy, usefulness, and impact variability responding to questions on COPD treatment. Given the generative intent of this project, the Delphi method was used to derive consensus rather than use a nominal group technique.[21] This is because the Delphi method explicitly forces convergence and ensures final ratings represent collective clinical judgement. Preliminary results were presented as a poster at the 2025 American College of Clinical Pharmacy Annual Meeting.[22] Reporting adhered to the ACcurate COnsensus Reporting Document guideline.[23] This study was exempt by the Institutional Review Board for Human Subjects at the University of Georgia, (in accordance with 45 CFR §46.104). We used ChatGPT-4o on 3 separate computers with distinct user accounts. Sessions occurred on May 6, 2025 with browsing enabled. Prompts were entered as the questions (verbatim) and are provided in the Appendix in the online supplement.

Five educational clinical questions were created by 3 residency-trained, board-certified pharmacists whose primary employment was a faculty position at a college of pharmacy. The pharmacists graduated and completed 2 years of postdoctoral training (postgraduate year-1 and -2 residency training) at different institutions, practiced in settings that differed in delivery of care (i.e., ambulatory care, inpatient), types of facilities (i.e., government, nonprofit), patient population (i.e., geriatrics, adults), and had a range of practice experience between 9 to 15 years; therefore, although the pharmacist panelists had similar expertise, they did not all practice at the same practice site

For personal use only. Permission required for all other uses.

(i.e., 3 distinct locations in Georgia and Oklahoma). All 3 pharmacists teach COPD pharmacotherapeutics within the didactic and experiential curriculum at their colleges.

The pharmacists used COPD as the primary focus of the questions, addressing domains such as evidence-based management, patient counseling considerations, and contraindications. The questions differed in format, ranging from well-defined structured questions to more open-ended ill-structured scenarios (Appendix in the online supplement). Well-structured questions typically provide key details explicitly and narrow the range of acceptable solutions, often mapping neatly onto established guidelines or a single best answer.[24] By comparison, ill-structured questions usually demand more nuanced clinical reasoning, requiring clinicians to synthesize patient-specific variables and weigh trade-offs among multiple defensible options to arrive at a prioritized, context-dependent decision. Each question was entered concurrently into 3 separate computers using ChatGPT-4o to generate responses (output). Notably, ChatGPT-4o is a subscription-based program that had internet access and knowledge cutoffs[18,25] ending in 2023. All generated responses were compiled and evaluated by (the same) 3 pharmacists using a modified Delphi process, as depicted in Figure 1.

Pharmacists independently assessed each response across 3 domains: accuracy, usefulness, and impact variability. Based on previous research on LLM, domains were separated on a 3-point Likert scale.[26,27] Accuracy reflected the degree to which the response aligned with correct or accepted clinical information (e.g., published guidelines, randomized controlled trials) and was ordinal scale scored as 0 (poor), 1 (borderline), or 2 (good). Usefulness represented an overall judgment of whether the response met the intended goal of the prompt either from a clinician or patient perspective depending on the prompt and was ordinal scale scored as 0 (not useful), 1 (somewhat useful), or 2 (very useful). Impact variability (sometimes referred to as either consistency or reliability) captured the extent to which differences in responses could meaningfully influence patient outcomes (e.g., clinical targets) and was ordinal scale scored as 0 (low), 1 (moderate), or 2 (high).[26,27] For impact variability, panelists judged whether differences among the 3 simultaneous responses could change medication management (e.g., inhaled therapy class, corticosteroid exposure, antibiotic use) or key clinical targets (e.g., exacerbation risk reduction). Operational definitions and examples for each level were predistributed to panelists (Appendix in the online supplement). Individual ratings were subsequently aggregated, anonymized, and redistributed to the pharmacists to allow comparison with group responses and the opportunity to make revisions or share additional commentary with the group. The study moderator summarized feedback and convened one virtual consensus meeting with the 3 pharmacists to address items with discordant ratings. Consensus was defined as unanimous agreement among all 3 pharmacists for each response and domain (i.e., accuracy, usefulness, impact variability) and achieved through discussion and agreement. Discordant ratings were reviewed item-by-item and rating changes were not decided by a majority vote. In some instances, discussion led panelists who were initially in agreement to revise toward the discordant rating. Prior to the consensus meeting, Fleiss' kappa was calculated for each domain across Delphi rounds 1 and 2 with 95% confidence intervals via bootstrapping to understand the reliability of each assessment[28,29]; we also summarized domain-specific kappa by question. Descriptive statistics were used to summarize the findings. Data analysis was performed in SAS version 9.4 (SAS Institute; Cary, North Carolina).

Results

Fleiss' kappa for Delphi rounds 1 and 2 are presented[30] in Table 1. Interrater reliability among the 3 panelists evaluating ChatGPT-4o's accuracy increased from slight agreement to fair agreement proceeding from Delphi round 1 to Delphi round 2. Poor agreement was observed regarding ChatGPT-4o's usefulness and impact variability in both Delphi rounds 1 and 2.

Among 5 COPD-related questions simultaneously administered across 3 instances of ChatGPT-4o, consensus among 3 board-certified pharmacists was achieved on 40 out of 45 responses (88.9% consensus agreement). Consensus was completely achieved regarding ChatGPT's accuracy though not achieved regarding the usefulness of one question and the impact variability of another question. The accuracy, usefulness, and impact variability across questions and responses are presented in Table 2.

ChatGPT-4o accuracy ranged from poor to good across the 5 questions. The consensus panelists agreed that ChatGPT-4o was borderline accurate on staging COPD, recommending initial inhaled medications, and assessing and planning a COPD medication treatment plan for a complex patient case. The consensus panel also felt that ChatGPT was borderline accurate responding to a question regarding treatment of an exacerbation of COPD (ECOPD) though one response was scored as good (versus others scored as borderline). The consensus panel agreed that ChatGPT-4o demonstrated poor accuracy responding to a question on selecting and recommending medications to treat stable COPD because each ChatGPT-4o response referenced an outdated clinical practice guideline.

ChatGPT-4o usefulness ranged between somewhat useful to very useful; the consensus panel did not score ChatGPT-4o as not useful across any responses (i.e., ChatGPT-4o was at least somewhat useful). The consensus panel felt that ChatGPT-4o was somewhat useful at staging

For personal use only. Permission required for all other uses.

Figure 1. Methodology Flow Diagram for Modified Delphi Technique

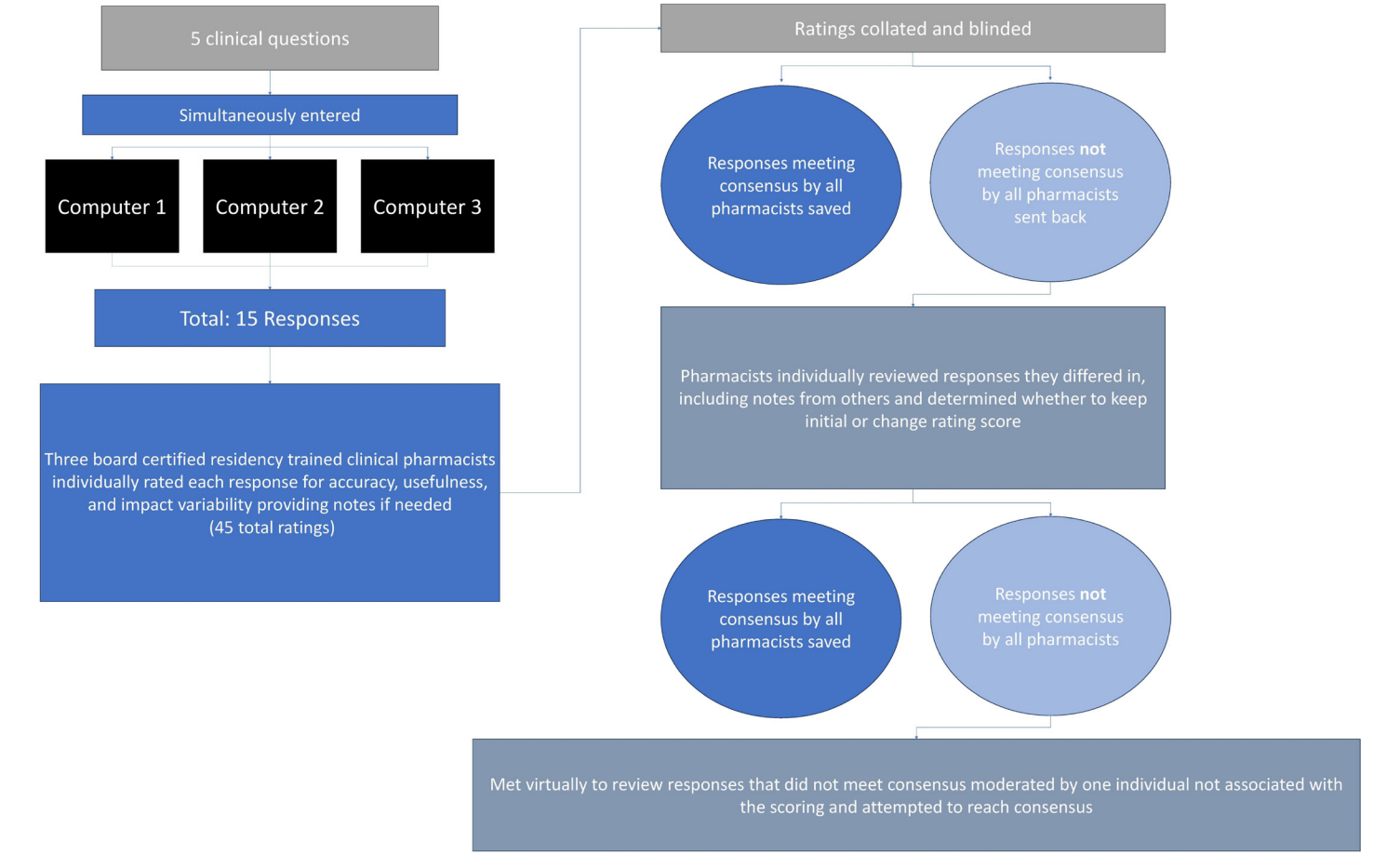

Table 1. Fleiss' Kappa by Delphi Round, Question, and Domain, with Bootstrapped 95% Confidence Intervals for Delphi Rounds 1 and 2

Delphi Round	Question	Accuracy κ (95% CI)	Usefulness κ (95% CI)	Impact Variability κ (95% CI)
1	**Overall**	0.22 (-0.08 to 0.50)	-0.02 (-0.26 to 0.17)	-0.02 (-0.27 to 0.16)
	1	0.10	-0.50	-0.13
	2	1.00	-0.29	-0.29
	3	-0.50	-0.35	-0.38
	4	0.00	0.36	-0.35
	5	-0.35	0.00	-0.04
2	**Overall**	0.43 (0.10 to 0.67)	-0.02 (-0.26 to 0.20)	-0.02 (-0.26 to 0.16)
	1	0.10	-0.50	-0.13
	2	1.00	-0.29	-0.29
	3	-0.50	-0.35	-0.38
	4	0.36	0.36	-0.35
	5	-0.13	0.00	-0.04

Fleiss' kappa was calculated across all responses within each Delphi round (N=15 responses per round, 3 panelists, 3-point ordinal scale). Each question-level kappa was calculated from 3 ChatGPT responses rated by 3 panelists on a 3-point ordinal scale. Due to the small number of participants per question (N=3), these estimates should be interpreted with caution.

CI=confidence interval

COPD and initiating medications and selecting medications to treat stable COPD. Regarding treating an ECOPD and assessing and planning a complex case, the consensus panel scored ChatGPT-4o as somewhat useful though one response (in each question) was scored as very useful. Consensus was not achieved among the panelists regarding ChatGPT's usefulness to a question asking to identify medication-related adverse events; herein, the consensus panel scored ChatGPT-4o somewhat to very useful.

ChatGPT-4o impact variability ranged between low to high. The consensus panel felt that ChatGPT-4o displayed high impact variability at staging COPD and initiating medication

For personal use only. Permission required for all other uses.

Table 2. Accuracy, Usefulness, and Impact Variability of ChatGPT Simultaneous Responses to COPD Questions

Question, Structure, and Topic	Accuracy[a]	Usefulness[b]	Impact Variability[c]
1; ill-structured;	Response 1: Borderline	Response 1: Somewhat	Response 1: High
GOLD staging and initial medication treatment	Response 2: Borderline	Response 2: Somewhat	Response 2: High
	Response 3: Borderline	Response 3: Somewhat	Response 3: High
2; structured;	Response 1: Good	Response 1: Very	Response 1: High
Medication-related adverse events	Response 2: Good	Response 2: Somewhat-to-Very[d]	Response 2: Moderate
	Response 3: Good	Response 3: Somewhat-to-Very[d]	Response 3: Moderate
3; ill-structured;	Response 1: Poor	Response 1: Somewhat	Response 1: Moderate[d]
Stable COPD medication treatment	Response 2: Poor	Response 2: Somewhat	Response 2: Low-to-Moderate[d]
	Response 3: Poor	Response 3: Somewhat	Response 3: Moderate[d]
4; ill-structured;	Response 1: Borderline	Response 1: Somewhat	Response 1: Moderate
ECOPD medication treatment	Response 2: Good	Response 2: Very	Response 2: Moderate
	Response 3: Borderline	Response 3: Somewhat	Response 3: Moderate
5; structured;	Response 1: Borderline	Response 1: Somewhat	Response 1: Low
Complex COPD case requiring assessment and plan	Response 2: Borderline	Response 2: Somewhat	Response 2: Moderate
	Response 3: Borderline	Response 3: Very	Response 3: Low

[a] Accuracy score scale: 0 (poor), 1 (borderline), 2 (good)
[b] Usefulness score scale: 0 (not useful), 1 (somewhat useful), 2 (very useful)
[c] Impact variability score scale: 0 (low), 1 (moderate), 2 (high)
[d] Did not reach consensus

COPD=chronic obstructive pulmonary disease; GOLD=Global initiative for chronic Obstructive Lung Disease; ECOPD=exacerbation of chronic obstructive pulmonary disease

and displayed moderate impact variability at treating an ECOPD. Regarding medication-related adverse events and complex case assessment and planning, the consensus panel scored ChatGPT-4o as moderate to high and low to moderate impact variability, respectively. Consensus was not achieved among the panelists regarding ChatGPT's impact variability to a question regarding selection of medications to treat stable COPD; notably, this was the same and only question for which the panelists agreed ChatGPT-4o displayed poor accuracy because it referenced an outdated clinical practice guideline.

Discussion

This appears to be one of the first studies evaluating simultaneous responses to clinical questions across 3 devices using the same LLM AI (i.e., ChatGPT-4o). Overall, some differences existed across accuracy, usefulness, and impact variability. Using a 3-round modified Delphi approach among residency-trained, board-certified pharmacist faculty, consensus was achieved on 89% of ChatGPT outputs following COPD prompts. The consensus panel of pharmacists agreed ChatGPT-4o accuracy ranged from poor to good, usefulness was at least somewhat useful and ranged between somewhat useful to very useful, and impact variability ranged between low and high. Though Fleiss' kappa measures of inter-rater reliability ranged from poor to fair during Delphi rounds 1 and 2, the pattern of improving agreement and the formation of consensus over time was expected given the use of a modified Delphi approach. Despite having an answer key and instructions, initial variability may reflect differences in the pharmacist panelists' clinical interpretation and reasoning. However, subsequent rounds allowed the panelists to review their colleagues' justification, promoting reflection and possible recalibration of their initial scores. The final consensus meeting provided an opportunity for structured discussion through the moderator to facilitate alignment. This is inherent to Delphi methods which typically results in improved inter-rater agreement over time.[31] Based on the literature, presenting these data are needed, since some studies propose that Delphi studies lack defining inter-rater reliability and should be included when possible.[32,33]

Though AI offers the promise of optimizing clinicians' workflow and addressing patients' clinical questions, concerns about AI's static knowledge base have been raised, particularly concerning medications.[11,17,18] A study by Imtiaz et al compared 2 LLM AIs, ChatGPT-3.5 and Bing, and scored ChatGPT-3.5 higher than Bing across accuracy, completeness, clarity, and safety domains among 21 COPD prompts.[14] Though question-level data were not provided, the authors noted that ChatGPT-3.5 referenced outdated clinical practice guidelines and provided incomplete responses with respect to inhaled medications including corticosteroids, long-acting beta2-agonists, and long-acting muscarinic antagonists. Results from our study identified similar accuracy concerns, with one response displaying poor accuracy concerning medication treatment for stable COPD because ChatGPT-4o referenced a retired copy of the Global initiative for chronic Obstructive Lung Disease (GOLD) Report.[34] These findings are consequential to health care providers and patients alike because international

For personal use only. Permission required for all other uses.

evidence-based guidelines for asthma and COPD, the Global Initiative for Asthma (GINA) and GOLD Reports, are updated annually.[35] ChatGPT's and other chatbots', including Bing and Claude, inability to capture, assess, and respond to contemporary and dynamic updates in pulmonary medicine are consequential and may lead to patients being either undertreated or overtreated with inhaled medications. Patients who are undertreated are at increased risk for a life-threatening ECOPD whereas patients who are overtreated are at increased risk for preventable adverse drug events. A plausible remedy to this problem includes routinely incorporating (uploading) clinical practice guidelines and landmark clinical trials into LLMs in real-time.[14] Future studies should evaluate if prompting, such as asking the model to use only guidelines published within a specified time-bound window, reduces outdated recommendations.[36]

Previous reports have criticized ChatGPT's ability to assess complex clinical cases.[17,25] In our project, ChatGPT's responses to a complex COPD case demonstrated borderline accuracy, somewhat usefulness, and moderate impact variability. The 2026 GOLD report strengthened its emphasis on avoiding unnecessary inhaled corticosteroid exposure (i.e., favoring dual inhaled long-acting bronchodilators unless there is clear eosinophilia or exacerbation-prone phenotype present) and clarified pharmacotherapy escalation after one ECOPD.[35] Despite moving away from the recommendation of using a combination of a long-acting beta2-agonist with an inhaled corticosteroid, without a long-acting muscarinic antagonist several years ago, clinicians continue to prescribe this combination inappropriately. A study conducted in Australia found that more than 50% of patients included were inappropriately prescribed an inhaled corticosteroid.[37] A cross-sectional study conducted within the U.S. Department of Veteran's Affairs, found that 23.9% of Veterans were prescribed an inhaled corticosteroid without an identifiable indication.[38] Even with guideline guidance on appropriate inhaled corticosteroid use and algorithms for transitioning patients from long-acting beta2-agonist/inhaled corticosteroid combination therapy to other more effective regimens, these errors persist.

Inappropriate COPD treatment extends beyond inhaled corticosteroids, too. A study conducted in the United States found that 45.3% of patients did not receive general guideline recommended inhaler treatment, further highlighting inappropriate use of medications in COPD.[39] Another facet to COPD management extends beyond stable patients to those presenting with exacerbations. A study of intensivists in France found that adherence to guidelines regarding the use of laboratory biomarkers and short-acting beta2-agonist medications were moderate and antibiotic prescribing was poor.[40]

Variability in acute COPD management is well-documented, with international guidelines from the European Respiratory Society and the American Thoracic Society highlighting strong recommendations and conditional areas precisely where LLM outputs must be cross-checked against current evidence-based guidance.[41] Our data, when combined with existing literature and clinical practice guidelines, shows that COPD management, specifically with inhaled corticosteroids, is complex and requires individualized patient assessment and clinical judgement. Because of this, LLM-generated recommendations with variable accuracy and usefulness can have a large impact on patient care. Prior to implementing these models for the management of COPD, the model should be trained by field experts, validated with updated clinical recommendations, and integrated with adequate safeguards to mitigate potentiation of medication errors. While this may be ideal, many clinicians using LLMs have not received formal training in prompt design, which may limit their ability to elicit accurate, clinically relevant, and appropriately tailored responses.[42,43] Therefore, our results may reflect the types of responses likely to occur during real-world use, where prompts are often developed without specialized prompt-engineering expertise. Collectively, results from our work and existing projects illustrate how either outdated or ambiguous advice, whether human or LLM-generated, can propagate COPD over- and undertreatment risks.

This project possesses limitations warranting discussion. Our data, when combined with existing literature, shows COPD management, specifically with inhaled corticosteroids, is complex and requires individualized patient assessment and clinical judgement. Clear guidance on Delphi panel composition is lacking; though only 3 board-certified pharmacists holding faculty appointments served as experts, increasing the panel size to include between 5 and 15 participants may have added robustness.[44] The small panel size also limited how much blinding could be preserved across multiple rounds. To mitigate this risk, discordant ratings were reviewed against operational definitions, and consensus was not treated as an issue of majority vote. Careful, moderated, item-level discussion allowed panelists to reconsider discordant ratings as systematically as possible though panel convergence towards consensus may have been susceptible to some degree of the bandwagon effect.[45,46] The use of a 3-point ordinal scale (0, 1, 2) to rate accuracy, usefulness, and impact variability may have centered consensus panelists towards the median response (1); rather, a 4-point ordinal scale (1, 2, 3, 4) could have been considered.[47] At the time of this project, ChatGPT-4o, only available via a paid subscription, was utilized; patients and providers using open-access (free) ChatGPT-3.5 may obtain different outputs, which have been shown to be less accurate than subscription-based versions.[18] Finally, we did not experimentally control computer temperature, a factor that plausibly influences inter-response variability and merits formal study.

For personal use only. Permission required for all other uses.

Conclusions

In this modified Delphi evaluation of simultaneous ChatGPT-4o responses to COPD medication management questions, variability was observed in accuracy and usefulness across responses generated at the same time using identical questions, and the differences identified may possess clinical implications. Although most responses were at least somewhat useful, the presence of outdated guideline references and differences in impact variability highlight important limitations for clinical applications. These findings suggest, at this time, LLMs should be considered supportive informational tools rather than stand-alone clinical decision-making resources. Given the evolving nature of evidence-based resources, careful clinician oversight remains essential to ensure safe, contemporary, and patient-centered care. Continued evaluation of AI performance, integration of updated clinical standards, and development of responsible implementation frameworks will be critical as AI is incorporated into health care delivery. Safe clinical use of LLMs in COPD care requires transparent grounding in current GOLD Report guidance, local oversight, and deployment patterns minimizing risks of outdated or inconsistent recommendations.

Acknowledgements

Author contributions: JC was responsible for conceptualization, project administration, resources, software, visualization, and supervision of the study. PMB and JC were responsible for data curation and PMB, DLL, RHS, and JC were responsible for formal analysis. JC and RP were responsible for methodology and PMB, DLL, RHS, and JC were responsible for study investigation. Validation was provided by PMB, RP, and JC. PMB, DLL, RHS, RP, RLL, KM, and JC were all responsible for writing, reviewing, and editing the manuscript and all authors reviewed and approved the final version of the manuscript submitted for publication.

Data availability statement: The data and analytic code underlying this study are available from the corresponding author upon reasonable request.

Declaration of Interest

The authors declare no conflicts of interest.

For personal use only. Permission required for all other uses.

References

1. World Health Organization (WHO). Chronic obstructive pulmonary disease (COPD). WHO website. Updated June 2026. Accessed February 2026. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

2. Adeloye D, Song P, Zhu Y, et al. Global, regional, and national prevalence of, and risk factors for, chronic obstructive pulmonary disease (COPD) in 2019: a systematic review and modelling analysis. *Lancet Respir Med.* 2022;10(5):447-458. https://doi.org/10.1016/S2213-2600(21)00511-7

3. Buist AS, McBurnie MA, Vollmer WM, et al. International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study. *Lancet.* 2007;370(9589):741-750. https://doi.org/10.1016/S0140-6736(07)61377-4

4. Ruvuna L, Sood A. Epidemiology of chronic obstructive pulmonary disease. *Clin Chest Med.* 2020;41(3):315-327. https://doi.org/10.1016/j.ccm.2020.05.002

5. Labaki WW, Rosenberg SR. Chronic obstructive pulmonary disease. *Ann Intern Med.* 2020;173(3):ITC17-ITC32. https://doi.org/10.7326/AITC202008040

6. Raherison C, Girodet PO. Epidemiology of COPD. *Eur Respir Rev.* 2009;18(114):213-221. https://doi.org/10.1183/09059180.00003609

7. Safiri S, Carson-Chahhoud K, Noori M, et al. Burden of chronic obstructive pulmonary disease and its attributable risk factors in 204 countries and territories, 1990-2019: results from the Global Burden of Disease Study 2019. *BMJ.* 2022;378:e069679. https://doi.org/10.1136/bmj-2021-069679

8. Boers E, Barrett M, Su JG, et al. Global burden of chronic obstructive pulmonary disease through 2050. *JAMA Netw Open.* 2023;6(12):e2346598. https://doi.org/10.1001/jamanetworkopen.2023.46598

9. Shatto JA, Stickland MK, Soril LJJ. Variations in COPD health care access and outcomes: a rapid review. *Chronic Obstr Pulm Dis.* 2024;11(2):229-246. https://doi.org/10.15326/jcopdf.2023.0441

10. Kaplan A, Cao H, FitzGerald JM, et al. Artificial intelligence/machine learning in respiratory medicine and potential role in asthma and COPD diagnosis. *J Allergy Clin Immunol Pract.* 2021;9(6):2255-2261. https://doi.org/10.1016/j.jaip.2021.02.014

11. Wong A, Flanagan T, Covington EW, et al. Forecasting the impact of artificial intelligence on clinical pharmacy practice. *J Am Coll Clin Pharm.* 2025;8(4):302-310. https://doi.org/10.1002/jac5.70004

12. Abdalla M, Saad M, Abazia D, et al. Responsible use of artificial intelligence in health care: evidence, challenges, and best practices: an opinion of the drug information practice and research network of the American College of Clinical Pharmacy. *J Am Coll Clin Pharm.* 2025;8(12):1333-1361. https://doi.org/10.1002/jac5.70131

13. Topalovic M, Das N, Burgel PR, et al. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *Eur Respir J.* 2019;53(4):1801660. https://doi.org/10.1183/13993003.01660-2018

14. 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(6):2400163. https://doi.org/10.1183/13993003.00163-2024

15. Merc P, Pirincci CS, Cihan E. Evaluation of AI chatbots for patient education and information on chronic obstructive pulmonary disease. *Heart Lung.* 2026;75:21-25. https://doi.org/10.1016/j.hrtlng.2025.09.002

16. Yin Y, Riaz Z, Sanchez RA, Mustafa A, Sedeh AE. Evaluating ChatGPT as a standalone tool for patient education: a review of frequently asked questions by patients with chronic obstructive pulmonary disease. *Cureus.* 2025;17(9):e92519. https://doi.org/10.7759/cureus.92519

17. Tran T, Le UM, Phan V. Assessing ChatGPT's response accuracy in pharmacy education: insights into cognitive complexity. *Am J Pharm Educ.* 2024;88(9):100943. https://doi.org/10.1016/j.ajpe.2024.100943

18. Khatri S, Sengul A, Moon J, Jackevicius CA. Accuracy and reproducibility of ChatGPT responses to real-world drug information questions. *J Am Coll Clin Pharm.* 2025;8(6):432-438. https://doi.org/10.1002/jac5.70038

19. Antao J, de Mast J, Marques A, Franssen FME, Spruit MA, Deng Q. Demystification of artificial intelligence for respiratory clinicians managing patients with obstructive lung diseases. *Expert Rev Respir Med.* 2023;17(12):1207-1219. https://doi.org/10.1080/17476348.2024.2302940

20. Mekov E, Miravitlles M, Petkov R. Artificial intelligence and machine learning in respiratory medicine. *Expert Rev Respir Med.* 2020;14(6):559-564. https://doi.org/10.1080/17476348.2020.1743181

21. Caudle KE, Burlison JD, Bourque MS, Hoffman JM. Research and scholarly methods: consensus techniques. *J Am Coll Clin Pharm.* 2025;8(7):610-618. https://doi.org/10.1002/jac5.70052

22. Caballero J, Lavender DL, Stone RH, et al. Assessing ChatGPT's accuracy and usefulness in medication management. In: 2025 ACCP annual meeting abstracts. *J Am Coll Clin Pharm.* 2025;8(12):1376-1503. https://doi.org/10.1002/jac5.70148

23. Gattrell WT, Logullo P, van Zuuren EJ, et al. ACCORD (ACcurate COnsensus Reporting Document): a reporting guideline for consensus methods in biomedicine developed via a modified Delphi. *PLoS Med.* 2024;21(1):e1004326. https://doi.org/10.1371/journal.pmed.1004326

24. Jonassen DH. *Learning to Solve Problems: A Handbook for Designing Problem-Solving Learning Environments.* Routledge; 2011.

For personal use only. Permission required for all other uses.

25. Tran BAC, Wantuch GA, Mangasar M, Miranda AC. Battle of the (chat) bots: comparative evaluation of April 2025 AI model accuracy on pharmacotherapeutic cases. *Am J Pharm Educ.* 2026;90(1):101922. https://doi.org/10.1016/j.ajpe.2025.101922

26. Ramasubramanian S, Balaji S, Kannan T, et al. Comparative evaluation of artificial intelligence systems' accuracy in providing medical drug dosages: a methodological study. *World J Methodol.* 2024;14(4):92802. https://doi.org/10.5662/wjm.v14.i4.92802

27. Varady NH, Lu AZ, Mazzucco M, et al. Understanding how ChatGPT may become a clinical administrative tool through an investigation on the ability to answer common patient questions concerning ulnar collateral ligament injuries. *Orthop J Sports Med.* 2024;12(7):23259671241257516. https://doi.org/10.1177/23259671241257516

28. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas.* 1973;33:613-619. https://doi.org/10.1177/001316447303300309

29. Haukoos JS, Lewis RJ. Advanced statistics: bootstrapping confidence intervals for statistics with "difficult" distributions. *Acad Emerg Med.* 2005;12(4):360-365. https://doi.org/10.1197/j.aem.2004.11.018

30. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-174. https://doi.org/10.2307/2529310

31. Hsu CC, Sandford BA. The Delphi technique: making sense of consensus. *Pract Assess Res. Eval.* 2007;12(1):10.

32. Niederberger M, Spranger J. Delphi technique in health sciences: a map. *Front Public Health.* 2020;8:457. https://doi.org/10.3389/fpubh.2020.00457

33. Spranger J, Homberg A, Sonnberger M, Niederberger M. Reporting guidelines for Delphi techniques in health sciences: a methodological review. *Z Evid Fortbild Qual Gesundhwes.* 2022;172:1-11. https://doi.org/10.1016/j.zefq.2022.04.025

34. Global Initiative for Obstructive Lung Disease (GOLD). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease, 2026 report. GOLD website. Published November 2025. Accessed March 2026. https://goldcopd.org/wp-content/uploads/2026/01/GOLD-REPORT-2026-v1.3-8Dec2025_WMV2.pdf

35. Global Initiative for Asthma (GINA). Global strategy for asthma management and prevention, 2025 GINA strategy report. GINA website. Published May 2025. Updated November 2025. Accessed March 2026. https://ginasthma.org/2025-gina-strategy-report/

36. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res.* 2023;25:e50638. https://doi.org/10.2196/50638

37. Harrison A, Borg B, Thompson B, Hew M, Dabscheck E. Inappropriate inhaled corticosteroid prescribing in chronic obstructive pulmonary disease patients. *Intern Med J.* 2017;47(11):1310-1313. https://doi.org/10.1111/imj.13611

38. Griffith MF, Feemster LC, Zeliadt SB, et al. Overuse and misuse of inhaled corticosteroids among veterans with COPD: a cross-sectional study evaluating targets for de-implementation. *J Gen Intern Med.* 2020;35:679-686. https://doi.org/10.1007/s11606-019-05461-1

39. Sharif R, Cuevas CR, Wang Y, Arora M, Sharma G. Guideline adherence in management of stable chronic obstructive pulmonary disease. *Respir Med.* 2013;107(7):1046-1052. https://doi.org/10.1016/j.rmed.2013.04.001

40. Haudebourg L, Faure M, Dres M, et al. Management of severe exacerbations of COPD by French intensivists and adherence to guidelines. *Respir Med Res.* 2025;87:101159. https://doi.org/10.1016/j.resmer.2025.101159

41. Wedzicha JA, Miravitlles M, Hurst JR, et al. Management of COPD exacerbations: a European Respiratory Society/American Thoracic Society guideline. *Eur Respir J.* 2017;49(3):1600791. https://doi.org/10.1183/13993003.00791-2016

42. Liu J, Liu F, Wang C, Liu S. Prompt engineering in clinical practice: tutorial for clinicians. *J Med Internet Res.* 2025;27:e72644. https://doi.org/10.2196/72644

43. Shah K, Xu AY, Sharma Y, et al. Large language model prompting techniques for advancement in clinical medicine. *J Clin Med.* 2024;13(17):5101. https://doi.org/10.3390/jcm13175101

44. Keeney S, Hasson F, McKenna H. *The Delphi Technique in Nursing and Health Research.* Wiley-Blackwell; 2011. https://doi.org/10.1002/9781444392029

45. O'Connor N, Clark S. Beware bandwagons! The bandwagon phenomenon in medicine, psychiatry and management. *Australas Psychiatry.* 2019;27(6):603-606. https://doi.org/10.1177/1039856219848829

46. Rikkers LF. The bandwagon effect. *J Gastrointest Surg.* 2002;6(6):787-794. https://doi.org/10.1016/S1091-255X(02)00054-9

47. Weinfurt KP, Reeve BB. Patient-reported outcome measures in clinical research. *JAMA.* 2022;328(5):472-473. https://doi.org/10.1001/jama.2022.11238

For personal use only. Permission required for all other uses.