

Review Articles



Large Language Models in Chronic Obstructive Pulmonary Disease Care and Health Management: A Scoping Review

Hanyu Du*

School of Public Health and Nursing, Hangzhou Normal University, Hangzhou, China

ARTICLE INFO

Article history:

Received 16 Sep. 2026

Accepted 15 Oct. 2026

Keywords:

Chronic obstructive pulmonary disease

Large language models

Patient education

Self-management

Pulmonary rehabilitation

Clinical decision support

ABSTRACT

Background: Chronic obstructive pulmonary disease (COPD) requires sustained patient education, self-management, medication use, pulmonary rehabilitation, symptom monitoring, and guideline-concordant clinical management. Large language models (LLMs) may expand access to conversational health information, but their value depends on more than factual accuracy. **Objective:** To map empirical evidence on LLMs and LLM-based generative artificial intelligence systems in COPD-related care and health management, summarize application domains and evaluation outcomes, and identify priorities for future research. **Methods:** A scoping review was conducted in accordance with Joanna Briggs Institute guidance and reported using PRISMA-ScR. PubMed/MEDLINE, Embase, Web of Science Core Collection, Scopus, CINAHL, and PsycINFO were searched from inception through September 10, 2026. Empirical studies evaluating an LLM or LLM-based generative AI system in COPD-related patient education, health information, pulmonary rehabilitation, medication management, or guideline-based clinical management were eligible. Data were charted and synthesized descriptively and thematically. **Results:** The search identified 327 records; 34 duplicates were removed, and 293 records underwent title/abstract screening. Of these, 281 were excluded, and 12 full-text reports were assessed. Four full-text reports were excluded, leaving eight studies. Five studies focused primarily on patient education or health information, one on pulmonary rehabilitation, one on COPD medication management, and one on guideline-based clinical management. Accuracy and information quality were the most common outcomes. LLMs often produced clinically acceptable responses, but limitations included incomplete information, high reading grade levels, outdated guideline references, response variability, and differences across models and task formats. Only two studies directly involved people with COPD, and one of these demonstrated a marked gap between favorable expert ratings and patients' ability to formulate disease-specific queries. **Conclusions:** Evidence on LLMs in COPD is expanding but remains dominated by model-response evaluation. Current findings support LLMs as supplementary, human-supervised tools rather than autonomous sources of COPD management advice. Future research should prioritize actionability, real-world usability, health literacy, equity, behavioral outcomes, clinical outcomes, and safety under routine patient use.

© Journal of Public Health and Preventive Medicine. Published by Public Health Young Scholars Alliance. All rights reserved.

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a major global public-health problem and requires long-term, multifaceted management. Beyond pharmacological treatment, effective COPD care depends on patients' ability to understand their condition, use treatments appropriately, recognize worsening symptoms, participate in pulmonary rehabilitation, maintain health-promoting behaviors, and seek professional care when

needed. These demands make access to understandable, current, and actionable information an important component of COPD management [1][2]. Large language models (LLMs) and LLM-based generative artificial intelligence systems can generate conversational responses to natural-language questions and are increasingly available to patients and clinicians. In COPD, potential applications include answering patient questions, supporting disease education, explaining pulmonary rehabilitation, assisting with medication-related decisions, and interpreting guideline-based management scenarios. The eight studies included in this review illustrate how quickly this evidence base has expanded across multiple

* Corresponding author: Hanyu Du. duhanyu1999@163.com

ISSN: 3071-4265/ This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: 10.64904/20261004



model families and use cases.

However, the clinical value of an LLM cannot be inferred from factual accuracy alone. Patient-facing information may be factually correct yet incomplete, difficult to read, insufficiently contextualized, or poorly actionable. In higher-risk areas such as medication management and exacerbation-related advice, outdated recommendations or inconsistent outputs may have direct clinical consequences. Moreover, model-level performance does not necessarily indicate that patients can formulate effective prompts, interpret answers correctly, or translate information into safe self-management behavior.

The emerging COPD literature is heterogeneous in models, question sources, evaluators, scoring systems, and outcomes. A scoping review is therefore appropriate to map the breadth of evidence, characterize how LLMs are currently being evaluated, and identify gaps that should shape future patient-centered and public-health research. This review aimed to map empirical evidence on LLMs in COPD care and health management; describe application domains, study designs, and evaluation outcomes; synthesize the principal findings; and identify priorities for future research and implementation.

2. Methods

2.1. Review design and reporting

This scoping review followed Joanna Briggs Institute methodological guidance for scoping reviews and was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR). The purpose was to map the available evidence rather than estimate a pooled intervention effect; therefore, meta-analysis was not planned [3][4].

2.2. Eligibility criteria

Eligibility was structured using the Population-Concept-Context

framework. The population/context of interest was COPD or COPD-related care. The concept was an LLM or LLM-based generative AI system used or evaluated for a COPD-related health-management task. Eligible contexts included patient education and health information, self-management support, pulmonary rehabilitation, medication management, and guideline-based clinical management. Studies involving clinician or expert evaluators were eligible when the evaluated task was directly relevant to COPD care or patient management.

Empirical studies were included if they (1) evaluated an LLM or LLM-based generative AI system; (2) addressed COPD as the primary disease context or reported separately identifiable COPD-specific analyses; (3) evaluated a COPD-related education, information, rehabilitation, medication-management, or clinical-management task; and (4) reported empirical performance, usability, or outcome data. Studies were excluded if they evaluated conventional machine learning or deep learning without an LLM component; focused on non-COPD populations without separable COPD findings; were purely technical natural-language-processing studies without a COPD care application; were editorials, commentaries, narrative discussions, or protocols without empirical outcome data; were duplicate/companion reports without distinct eligible evidence; or fell outside the prespecified health-management scope.

2.3. Information sources and search strategy

PubMed/MEDLINE, Embase, Web of Science Core Collection, Scopus, CINAHL, and PsycINFO were searched from inception through September 10, 2026. The strategy combined COPD-related terms with generic LLM/generative-AI terminology and names of commonly used systems. Because terminology in this rapidly evolving field is inconsistent, the primary strategy was deliberately broad and did not require a third health-management concept block. The complete database-specific strategies are provided in Supplementary Table S1.

Table S1. Database search strategies

Database	Search strategy
PubMed/MEDLINE	("Pulmonary Disease, Chronic Obstructive"[Mesh] OR COPD[tiab] OR "chronic obstructive pulmonary disease"[tiab] OR "chronic obstructive lung disease"[tiab] OR emphysema[tiab] OR "chronic bronchitis"[tiab]) AND ("large language model"[tiab] OR LLM*[tiab] OR "generative artificial intelligence"[tiab] OR "generative AI"[tiab] OR ChatGPT[tiab] OR GPT[tiab] OR "GPT-3"[tiab] OR "GPT-3.5"[tiab] OR "GPT-4"[tiab] OR "GPT-4o"[tiab] OR "GPT-5"[tiab] OR Gemini[tiab] OR Bard[tiab] OR Claude[tiab] OR DeepSeek[tiab] OR Copilot[tiab] OR Llama[tiab] OR Qwen[tiab] OR "conversational AI"[tiab] OR chatbot*[tiab])
Embase (Ovid syntax)	1 exp chronic obstructive pulmonary disease/ 2 (COPD or chronic obstructive pulmonary disease or chronic obstructive lung disease or emphysema or chronic bronchitis).ti,ab,kw. 3 1 or 2 4 exp artificial intelligence/ 5 exp natural language processing/ 6 (large language model* or LLM* or generative artificial intelligence or generative AI or ChatGPT or GPT-3 or GPT-3.5 or GPT-4 or GPT-4o or GPT-5 or Gemini or Bard or Claude or DeepSeek or Copilot or Llama or Qwen or conversational AI or chatbot*).ti,ab,kw. 7 4 or 5 or 6 8 3 and 7
Web of Science Core Collection	TS=(COPD OR "chronic obstructive pulmonary disease" OR "chronic obstructive lung disease" OR emphysema OR "chronic bronchitis") AND TS=("large language model*" OR LLM* OR "generative artificial intelligence" OR "generative AI" OR ChatGPT OR GPT OR "GPT-3" OR "GPT-3.5" OR "GPT-4" OR "GPT-4o" OR "GPT-5" OR Gemini OR Bard OR Claude OR DeepSeek OR Copilot OR Llama OR Qwen OR "conversational AI" OR chatbot*)
Scopus	TITLE-ABS-KEY(COPD OR "chronic obstructive pulmonary disease" OR "chronic obstructive lung disease" OR emphysema OR "chronic bronchitis") AND TITLE-ABS-KEY("large language model*" OR LLM* OR "generative artificial intelligence" OR "generative AI" OR ChatGPT OR GPT OR "GPT-3" OR "GPT-3.5" OR "GPT-4" OR "GPT-4o" OR "GPT-5" OR Gemini OR Bard OR Claude OR DeepSeek OR Copilot OR Llama OR Qwen OR "conversational AI" OR chatbot*)
CINAHL (EBSCOhost syntax)	((MH "Pulmonary Disease, Chronic Obstructive+") OR TI (COPD OR "chronic obstructive pulmonary disease" OR "chronic obstructive lung disease" OR emphysema OR "chronic bronchitis") OR AB (COPD OR "chronic obstructive pulmonary disease" OR "chronic obstructive lung disease" OR emphysema OR "chronic bronchitis")) AND ((MH "Artificial Intelligence+") OR (MH "Natural Language Processing") OR TI ("large language model*" OR LLM* OR "generative artificial intelligence" OR "generative

Database	Search strategy
PsycINFO (EBSCOhost syntax)	AI" OR ChatGPT OR GPT OR "GPT-3" OR "GPT-3.5" OR "GPT-4" OR "GPT-4o" OR "GPT-5" OR Gemini OR Bard OR Claude OR DeepSeek OR Copilot OR Llama OR Qwen OR "conversational AI" OR chatbot*) OR AB ("large language model*" OR LLM* OR "generative artificial intelligence" OR "generative AI" OR ChatGPT OR GPT OR "GPT-3" OR "GPT-3.5" OR "GPT-4" OR "GPT-4o" OR "GPT-5" OR Gemini OR Bard OR Claude OR DeepSeek OR Copilot OR Llama OR Qwen OR "conversational AI" OR chatbot*)) (DE "Chronic Obstructive Pulmonary Disease" OR TI (COPD OR "chronic obstructive pulmonary disease" OR "chronic obstructive lung disease" OR emphysema OR "chronic bronchitis")) OR AB (COPD OR "chronic obstructive pulmonary disease" OR "chronic obstructive lung disease" OR emphysema OR "chronic bronchitis")) AND (DE "Artificial Intelligence" OR DE "Natural Language Processing" OR TI ("large language model*" OR LLM* OR "generative artificial intelligence" OR "generative AI" OR ChatGPT OR GPT OR "GPT-3" OR "GPT-3.5" OR "GPT-4" OR "GPT-4o" OR "GPT-5" OR Gemini OR Bard OR Claude OR DeepSeek OR Copilot OR Llama OR Qwen OR "conversational AI" OR chatbot*) OR AB ("large language model*" OR LLM* OR "generative artificial intelligence" OR "generative AI" OR ChatGPT OR GPT OR "GPT-3" OR "GPT-3.5" OR "GPT-4" OR "GPT-4o" OR "GPT-5" OR Gemini OR Bard OR Claude OR DeepSeek OR Copilot OR Llama OR Qwen OR "conversational AI" OR chatbot*))

2.4. Study selection

All retrieved records were deduplicated before title and abstract screening. The review author screened titles and abstracts and subsequently assessed the full text of potentially eligible reports against the prespecified eligibility criteria. Screening decisions and reasons for full-text exclusion were documented during the review process. The database search identified 327

records. After removal of 34 duplicates, 293 records remained for screening. Of these, 281 were excluded at title/abstract screening. Twelve reports were assessed in full text; four were excluded after full-text review: one duplicate/companion report, one study protocol without empirical outcome data, one report with a non-COPD primary focus, and one study outside the prespecified health-management scope. Eight studies were therefore included. The study-selection process is shown in Figure 1.

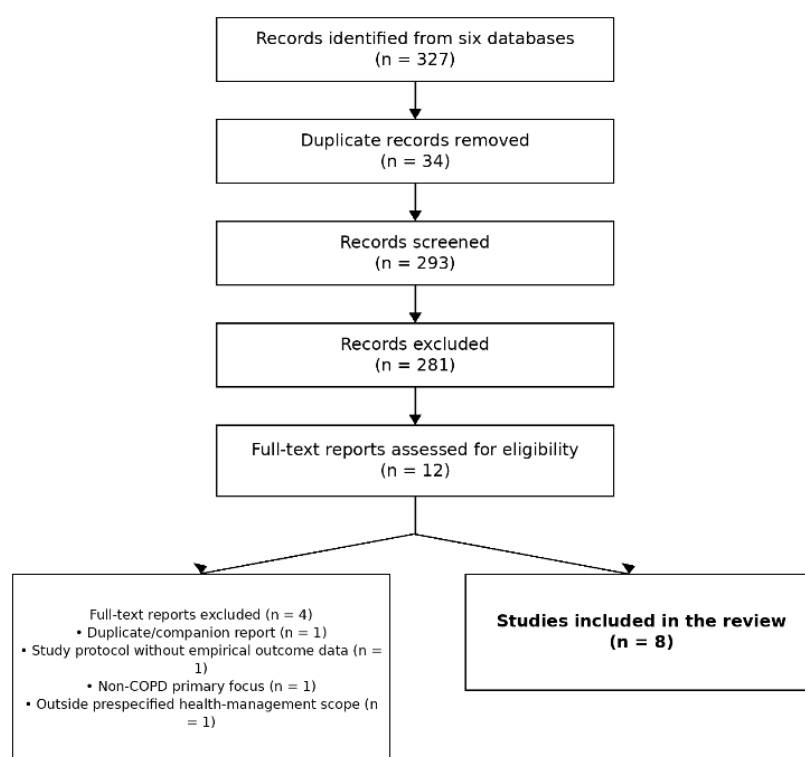


Figure 1. Study-selection flow diagram.

Note. Full-text exclusions: duplicate/companion report (n=1); study protocol without empirical outcome data (n=1); non-COPD primary focus (n=1); outside the prespecified health-management scope (n=1).

2.5. Data charting

A standardized charting framework was used to extract publication year,

country or study setting, study design, LLM(s) evaluated, source and number of questions/prompts, evaluator characteristics, direct patient involvement, evaluation domains, and principal findings. Particular attention was paid to accuracy, comprehensiveness or scientific adequacy,

readability or understandability, reliability or reproducibility, safety, usefulness, response variability, and real-world patient use.

2.6. Synthesis

Owing to substantial heterogeneity in study designs, models, prompts, scoring systems, and outcome definitions, findings were synthesized descriptively and thematically rather than statistically pooled. Studies were grouped by dominant application domain: patient education/health information, pulmonary rehabilitation, medication management, and guideline-based clinical management. No formal risk-of-bias assessment was used to determine inclusion because the objective of this scoping review was evidence mapping rather than comparative effect estimation.

3. Results

3.1. Study selection and overall characteristics

Eight studies met the inclusion criteria. The included evidence was published in 2025–2026 and represented a rapidly developing field. Study settings included international collaborations, South Korea, Türkiye, China, and the United States. Five studies primarily evaluated patient-facing education or health-information responses, one focused on pulmonary rehabilitation, one on medication management, and one on guideline-based COPD management. Most studies were cross-sectional or benchmark-style evaluations of generated text. Only two studies directly involved people with COPD: the AIR-COPD Study included 13 patient representatives in understandability assessment and question generation, whereas Yi et al. included 10 patients in an information-seeking task.

The evaluated systems included ChatGPT/GPT models, Bard, Copilot, Gemini, DeepSeek, Perplexity, Grok, Manus, and Kimi. Question sets ranged from five medication-management scenarios to 90 GOLD-derived questions. Evaluation was usually performed by pulmonologists, physiotherapists, pharmacists, or multidisciplinary experts, with substantial variation in scoring methods and outcome definitions [5][12].

3.2. Patient education and health information

Five studies primarily evaluated patient education or health information. In the AIR-COPD Study, 15 questions formulated by COPD patient representatives were submitted to ChatGPT-3.5, Bard, and Copilot. Twenty clinical experts assessed accuracy and comprehensiveness, and 13 patient representatives assessed understandability. ChatGPT was the most reliable (14/15 responses) and had higher median accuracy (8.0, IQR 7.0–9.0) and comprehensiveness (8.0, IQR 6.8–9.0) than Bard and Copilot, whereas understandability did not differ significantly across models [5].

Merç et al. evaluated ChatGPT-4 Turbo, Gemini 2.0 Flash, and DeepSeek R1 using 10 commonly asked COPD questions. Five physiotherapists rated response quality, and readability was assessed with the Flesch-Kincaid Grade Level (FKGL). Overall response quality was similar for most questions, although Gemini scored better on one question. Readability varied materially: DeepSeek generally produced lower reading-grade requirements, whereas Gemini frequently generated more complex, academic-level text [7].

Yin et al. submitted 44 patient-centered COPD questions to ChatGPT-4o three times using separate accounts, producing 132 responses. Only 20.5% (9/44) of responses were fully accurate, while 79.5% (35/44) were partially accurate because they omitted information or nuance; no response was

classified as inaccurate. Reproducibility across three iterations was 93.2% (41/44), but inter-rater agreement between the two primary physicians was only moderate (Cohen kappa 0.62), underscoring the subjectivity involved in distinguishing complete from partially complete answers [8].

Güçsav et al. compared ChatGPT-4o and Gemini 2.5 Flash across 34 common COPD questions. Four pulmonologists blindly evaluated accuracy, clarity, and scientific adequacy. Both models were generally rated favorably. Gemini had higher overall scientific adequacy (4.69 ± 0.31 vs 4.34 ± 0.45 ; $P < 0.001$), while differences in accuracy and clarity were not significant overall. Gemini generated substantially longer responses, making the scientific-adequacy advantage difficult to separate from response-length effects [9].

Chen et al. compared ChatGPT-4.0, DeepSeek, Gemini, and Perplexity using 20 English-language questions covering basic knowledge, nursing measures, rehabilitation, and prevention. Three COPD experts rated 80 responses. Accuracy and comprehensiveness differed significantly across models; ChatGPT-4.0 and DeepSeek performed similarly, and both outperformed Gemini and Perplexity. No model consistently achieved an approximately eighth-grade reading level, and Perplexity had the highest FKGL. The study was explicitly designed for clinician-assisted rather than autonomous patient education [12].

3.3. Pulmonary rehabilitation and patient usability

Yi et al. developed 19 pulmonary-rehabilitation questions through a two-round Delphi process and evaluated ChatGPT-4o using 20 multidisciplinary experts. Mean scores on a 4-point scale were high for accuracy (3.63), readability (3.57), safety (3.68), and overall quality (3.64), and semantic similarity across duplicate generations was high when assessed using an embedding model (mean 0.94, SD 0.03). However, objective readability remained demanding: mean FKGL was 13.49 and mean SMOG was 14.56. In the separate patient task, only 2 of 10 patients successfully retrieved COPD-specific pulmonary-rehabilitation information after appropriately contextualizing their condition. Most patients did not initially disclose disease-specific context, and lower digital literacy was associated with difficulty using the system. This study provided the clearest evidence of a gap between expert-rated response quality and real-world patient use [6].

3.4. Medication management

Boylan et al. used a three-round modified Delphi approach to evaluate ChatGPT-4o responses to five COPD medication-management questions. Each question was entered simultaneously on three separate computers, generating 15 responses that were rated by three residency-trained, board-certified clinical pharmacists for accuracy, usefulness, and impact variability. Consensus was achieved for 40 of 45 response-domain ratings (88.9%). Accuracy ranged from poor to good and usefulness from somewhat to very useful. Importantly, all three responses to the stable-COPD pharmacotherapy question cited a retired clinical-practice guideline, resulting in poor accuracy ratings. Simultaneous outputs also differed in ways that could alter medication decisions, supporting the need for current guideline grounding and local clinical oversight [11].

3.5. Guideline-based clinical management

Al-Waqeerah et al. evaluated ChatGPT-5, DeepSeek-V3.2, Gemini 3, Grok, Manus, and Kimi K1.5 using 90 questions derived from the 2025 GOLD

report: 30 yes/no, 30 multiple-choice, and 30 open-ended questions. All questions were administered daily for seven days, yielding 3,780 responses. Binary accuracy was highest for Kimi K1.5 (97.1%, 408/420) and lowest for Gemini 3 (89.5%, 376/420). In contrast, Gemini 3 achieved the highest weighted score for open-ended clinical reasoning (96.5%) and DeepSeek-V3.2 the lowest (91.8%). All models were statistically stable across the seven days. The inversion between binary and open-ended performance demonstrates that apparent model ranking depends strongly on task format and outcome definition^[10].

3.6. Cross-study patterns in evaluation outcomes

Across studies, accuracy was the most consistently evaluated construct, but

definitions and scoring systems differed markedly. Completeness, comprehensiveness, scientific adequacy, and quality were assessed in several studies, whereas direct actionability was rarely measured. Readability was assessed in three studies and repeatedly emerged as a limitation: objective reading levels were frequently above common patient-education targets even when clinicians rated responses as clear or readable. Reliability and reproducibility were evaluated through repeated prompts, simultaneous devices, or longitudinal testing, with generally favorable stability in some studies but clinically meaningful variation in others. Safety was assessed directly in the pulmonary-rehabilitation study and indirectly through expert judgments or detection of outdated guidance in several others. Direct patient usability and behavioral or clinical outcomes were largely absent.

Table 1. Characteristics of included studies

Study	Setting/design	LLM(s)	Questions/prompts	Evaluators/patient involvement	Main outcomes
Nigro et al. (AIR-COPD), 2026	International cross-sectional comparative study; questions submitted in Milan, Italy	ChatGPT-3.5; Bard; Copilot	15 patient-generated questions; easy/medium/difficult	20 clinical experts; 13 COPD patient representatives; blinded	Reliability; accuracy; comprehensiveness; understandability
Yi et al., 2026	South Korea; Delphi-based expert evaluation plus exploratory patient task	ChatGPT-4o	19 PR questions generated after 2-round Delphi; responses generated twice	20 multidisciplinary experts; 10 patients with COPD	Accuracy; readability; safety; overall quality; reproducibility; patient information retrieval
Merç et al., 2026	Türkiye; observational cross-sectional comparative study	ChatGPT-4 Turbo; Gemini 2.0 Flash; DeepSeek R1	10 commonly asked COPD questions	5 physiotherapists; no direct patient testing	Response quality; FKGL readability
Yin et al., 2025	Non-human-subject observational evaluation; international authors, U.S.-based physician raters	ChatGPT-4o	44 questions × 3 accounts = 132 responses	2 COPD-experienced physicians; third reviewer for disagreements; no patients	Accuracy; reproducibility; inter-rater reliability
Güçsav et al., 2026	Türkiye; comparative blinded expert-panel study	ChatGPT-4o; Gemini 2.5 Flash	34 public COPD questions	4 pulmonologists; no patients	Accuracy; clarity; scientific adequacy; response length
Al-Waqeerah et al., 2026	China; 7-day longitudinal benchmark against GOLD 2025	ChatGPT-5; DeepSeek-V3.2; Gemini 3; Grok; Manus; Kimi K1.5	90 GOLD-derived questions daily for 7 days; 3,780 responses	Guideline-based scoring; clinician review/audit of open-ended responses	Binary accuracy; open-ended reasoning quality; consistency
Boylan et al., 2026	United States; 3-round modified Delphi	ChatGPT-4o	5 medication-management questions × 3 devices = 15 responses	3 board-certified clinical pharmacists; no patients	Accuracy; usefulness; impact variability
Chen et al., 2026	China; comparative clinician-assisted patient-education evaluation	ChatGPT-4.0; DeepSeek; Gemini; Perplexity	20 questions × 4 LLMs = 80 responses	3 COPD experts; no patients	Accuracy; comprehensiveness; readability

Table 2. Principal findings of included studies

Study	Key quantitative/qualitative findings	Interpretation for COPD care
AIR-COPD	ChatGPT reliability 14/15; median accuracy 8.0 and comprehensiveness 8.0; understandability similar across models (P=0.53).	Patient-generated questions can be answered reasonably well, but model differences are outcome-specific.
Yi et al.	Expert scores >3.5/4 across domains; embedding similarity 0.94; mean FKGL 13.49; only 2/10 patients retrieved COPD-specific PR information.	Strong expert ratings did not guarantee real-world usability; digital literacy and prompt contextualization matter.
Merç et al.	Overall quality similar for most questions; Gemini superior on one item; DeepSeek generally lowest reading difficulty.	Readability differs substantially across models even when information quality is comparable.
Yin et al.	20.5% fully accurate; 79.5% partially accurate; 0% inaccurate; reproducibility 93.2%; kappa 0.62.	Omission and incomplete nuance may be more common than outright falsehoods.
Güçsav et al.	Gemini scientific adequacy 4.69±0.31 vs ChatGPT 4.34±0.45 (P<0.001); no overall accuracy/clarity difference; Gemini responses longer.	Model comparisons can be confounded by response length and communication style.
Al-Waqeerah et al.	Binary accuracy 89.5%-97.1%; open-ended weighted scores 91.8%-96.5%; no significant temporal instability across 7 days.	Performance is high but task-format dependent; residual error persists even with updated models.
Boylan et al.	40/45 ratings reached consensus; one stable-COPD medication question was poorly accurate because all outputs cited a retired guideline.	Medication advice is a high-risk use case requiring current guideline grounding and clinician oversight.
Chen et al.	Accuracy and comprehensiveness differed across models (both P<0.001); ChatGPT-4.0 and DeepSeek outperformed Gemini and Perplexity; no model consistently met ~8th-grade readability target.	Clinician review remains necessary before patient-facing use; readability is a persistent barrier.

4. Discussion

4.1. Principal findings

This scoping review identified eight empirical studies evaluating LLMs in COPD-related care and health management. The literature is recent, technically heterogeneous, and concentrated on generated-response quality rather than patient outcomes. Patient education and health information dominated the evidence base, while pulmonary rehabilitation, medication management, and guideline-based clinical reasoning were represented by single studies. Across applications, the evidence supports cautious optimism: LLMs often generated clinically acceptable information, but performance varied by model, task, evaluation criterion, and prompt context.

A central finding is that factual correctness alone is an insufficient endpoint. Yin et al. observed no clearly inaccurate responses, yet nearly four-fifths were only partially accurate because important information or nuance was missing. The AIR-COPD Study similarly separated accuracy from comprehensiveness and understandability. Chen et al. showed that accuracy, comprehensiveness, and readability can diverge, while Güçsav et al. demonstrated that longer answers can score higher for scientific adequacy without necessarily representing a better patient-facing response. Together, these findings argue for multidimensional evaluation rather than a single global quality score [8][9].

4.2. From model capability to patient action

The most important evidence gap concerns translation from technically acceptable output to usable patient support. Only two of the eight studies directly involved people with COPD. The AIR-COPD Study incorporated the patient voice in question generation and understandability ratings, but did not evaluate subsequent behavior. Yi et al. moved further toward real-world use and found that only 2 of 10 patients retrieved disease-specific pulmonary-rehabilitation information after appropriately contextualizing their condition, despite favorable expert ratings of ChatGPT responses. This discrepancy indicates that interaction skills, digital literacy, and prompt formulation are part of the intervention, not merely user characteristics external to it [5][6].

A useful conceptual pathway for future work is: LLM capability → communication quality → patient understanding → actionability → behavior change → clinical outcome → population-health impact. Current COPD evidence is concentrated on the first two stages. Studies rarely assessed whether information changed inhaler behavior, medication adherence, rehabilitation participation, exacerbation recognition, care-seeking, quality of life, or healthcare utilization. Without these downstream outcomes, high model scores should not be interpreted as evidence of clinical effectiveness.

4.3. Readability, health literacy, and equity

Readability emerged as a persistent weakness. In Yi et al., objective reading levels were substantially higher than typical patient-education targets despite strong expert-rated readability. Merç et al. showed meaningful differences among models, and Chen et al. found that none of four LLMs consistently met an approximate eighth-grade target. These findings reinforce the distinction between clinician-perceived clarity and patient-

accessible language [6][7][12].

The COPD population includes many older adults and people with varying levels of health and digital literacy. Yi et al. directly observed difficulty among older patients with limited prior exposure to ChatGPT, while Chen et al. emphasized the importance of language, culture, and health literacy for generalizability. If LLM-based education requires sophisticated prompt construction or high reading ability, implementation could widen rather than reduce disparities. Future evaluations should therefore stratify usability and outcomes by age, education, health literacy, digital literacy, language, and access to digital technology [1][6][12].

4.4. Safety, guideline currency, and human oversight

Safety concerns become more consequential as LLM use moves from general education toward treatment decisions. Boylan et al. showed that simultaneous ChatGPT-4o responses to identical COPD medication questions could differ materially and that all three outputs for one stable-COPD pharmacotherapy scenario relied on a retired guideline. Al-Waqeerah et al. demonstrated high adherence to the 2025 GOLD report across six contemporary systems, but residual error remained, and model rankings changed according to whether questions were binary or open-ended. These findings indicate that current models may be useful for information support but should not be treated as self-validating clinical decision systems [10][11].

For higher-risk COPD tasks, implementation should include explicit grounding in current guidelines, version control, documentation of model and account settings, mechanisms for escalation to clinicians, and local oversight. Patient-facing systems should clearly communicate uncertainty and avoid presenting individualized medication changes, oxygen adjustments, or exacerbation management decisions as definitive advice without clinical review.

4.5. Implications for nursing and public health

For nursing and public health practice, LLMs may be most immediately useful as clinician-assisted drafting and education tools. They can rapidly generate explanations, self-management summaries, or responses to common questions, after which nurses or other qualified clinicians can verify content, simplify language, add local care pathways, and tailor recommendations to the individual patient. This use case aligns with the clinician-assisted design explicitly evaluated by Chen et al. and with the repeated recommendation for professional validation across the included studies [6][8][11][12].

At a population level, conversational AI could increase the reach of COPD education where consultation time, specialist access, or continuity of care is limited. Yet scale should not be equated with equity. Public-health evaluation should include not only access but also who can successfully use the technology, whether information is understandable across literacy levels and languages, whether unsafe advice is recognized, and whether deployment changes healthcare utilization or disparities.

4.6. Priorities for future research

Future studies should progress from static response benchmarking toward prospective, patient-centered evaluation. Priority outcomes include comprehension, actionability, task completion, inhaler technique, medication adherence, pulmonary-rehabilitation participation, symptom

monitoring, exacerbation recognition, appropriate care-seeking, quality of life, and healthcare utilization. Randomized or pragmatic comparative studies are needed to determine whether LLM-assisted education adds benefit beyond standard education or conventional online information. Methodological reporting also requires greater standardization. Studies should record the exact model/version, access tier, browsing or retrieval settings, prompt wording, date/time of querying, whether memory was enabled, number of repeated generations, and the reference standard used for evaluation. Evaluation frameworks should distinguish accuracy from completeness, readability, understandability, actionability, safety, and reproducibility. Direct patient participation should become routine rather than exceptional.

4.7. Strengths and limitations of this review

This review has several strengths. It searched six major databases, used a broad LLM-focused strategy suitable for rapidly changing terminology, and restricted inclusion to empirical COPD-specific evaluations. The synthesis also distinguished model-level accuracy from patient-level usability and incorporated studies spanning education, rehabilitation, medication management, and guideline-based care.

The review also has limitations. First, only eight studies were eligible, and the literature is expanding rapidly. Second, the included studies used heterogeneous models, prompts, reference standards, scoring systems, and evaluator groups, limiting direct comparison. Third, most studies evaluated English-language prompts and generated text rather than real-world patient outcomes. Fourth, model behavior can change after software updates, making findings time-sensitive. Fifth, direct patient participation was limited to two studies, restricting conclusions about usability, actionability, and equity. Sixth, study selection and data charting were conducted by one reviewer, which may increase the risk of selection or extraction error compared with duplicate independent review. Finally, the review did not perform a formal risk-of-bias appraisal, consistent with its evidence-mapping objective; therefore, the findings should be interpreted as a characterization of the available evidence rather than a comparative ranking of study quality.

5. Conclusion

Empirical evidence on LLMs in COPD care is emerging rapidly but remains centered on model-response evaluation. Across eight studies, LLMs demonstrated promising performance in patient education, pulmonary rehabilitation information, medication management, and guideline-based clinical tasks, yet limitations were evident in completeness, reading level, guideline currency, response variability, and patient usability. The available evidence supports LLMs as supplementary, human-supervised tools rather than autonomous sources of COPD management advice. The next phase of research should move beyond the question of whether an LLM can produce a correct answer and determine whether it can deliver understandable, actionable, safe, equitable support that improves patient behavior and clinical outcomes.

Declaration of generative AI and AI-assisted technologies in the writing process

The free versions of Grammarly were used for spell-checking and grammar corrections. After using these tools, the author(s) reviewed and edited the

content as needed and took full responsibility for the content of the publication.

Funding

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

Declaration of competing interests

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data analyzed in this scoping review are derived from the published studies cited in the manuscript.

REFERENCES

- [1] World Health Organization. Chronic obstructive pulmonary disease (COPD). WHO; 2026.
- [2] Global Initiative for Chronic Obstructive Lung Disease. Global Strategy for Prevention, Diagnosis and Management of COPD: 2026 Report. GOLD; 2026.
- [3] Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. *JBIM Evid Synth*. 2020;18(10):2119–2126. doi:10.1112/JBIES-20-00167.
- [4] Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169(7):467–473. doi:10.7326/M18-0850.
- [5] Nigro M, Behring GE, Aliverti A, et al. Accuracy, comprehensiveness and understandability of AI-generated answers to questions from people with COPD: the AIR-COPD Study. *Respir Res*. 2026;27:19. doi:10.1186/s12931-025-03438-9.
- [6] Yi YG, Lee M, Heo J, et al. ChatGPT for pulmonary rehabilitation in chronic obstructive pulmonary disease is clinically credible but challenging for patients to use. *Sci Rep*. 2026. doi:10.1038/s41598-026-66051-w.
- [7] Merç P, Şahbaz Pirinççi C, Cihan E. Evaluation of AI chatbots for patient education and information on chronic obstructive pulmonary disease. *Heart Lung*. 2026;75:21–25. doi:10.1016/j.hrtlng.2025.09.002.
- [8] Yin Y, Riaz Z, Amoro Sanchez R, Mustafa A, Eighaei Sedeh A. 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. doi:10.7759/cureus.92519.
- [9] Güçsav MO, Serçe Unat D, Akçay O, Unat ÖS, Ayrancı A, Erbaycu AE. Comparative assessment of ChatGPT and Gemini answers to common chronic obstructive pulmonary disease questions: an expert panel evaluation by pulmonologists. *Chronic Respir Dis*. 2026;23:1–10. doi:10.1177/14799731261443321.
- [10] Al-waqeerah A, Bashah A, Ghaleb E, He H, Gao L. AI in respiratory care: findings from the GOLD report. *J Transl Med*. 2026;24:1082. doi:10.1186/s12967-026-08454-x.
- [11] 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:10.15326/jcopdf.2026.0794.
- [12] Chen Z, Jiang Y, Jiang H, Xie H, Wu L, Shen Y. Comparative performance of ChatGPT-4.0, DeepSeek, Gemini and Perplexity in answering common questions from patients with COPD. *Sci Rep*. 2026. doi:10.1038/s41598-026-70606-2.