Acta Medica Young Doctors
ISSN: 3108-4532
Volume 2 Issue 4

Post-Discharge Advice After Spontaneous Pneumothorax from Freely Accessible Consumer AI Chatbots: A Scenario-Based Study of Guideline Concordance and Clinical Safety

Original Article

Okan Karataş[1]

1. Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye

ABSTRACT

Aim: Large language models (LLMs) are increasingly accessible to patients seeking health-related information, including practical advice after treatment for spontaneous pneumothorax. The clinical usefulness of such responses depends not only on overall guideline concordance but also on the avoidance of potentially unsafe recommendations. In the present study, the evaluated advice reflected the models assigned by default through each provider's free consumer interface rather than investigator-selected model configurations.

Methodology: Thirty standardized synthetic scenarios representing routine discharge, lifestyle and travel, recurrence and high-risk situations, secondary spontaneous pneumothorax, and post-discharge emergencies were evaluated using three freely accessible consumer-facing LLMs: ChatGPT (GPT-5.6 Luna), Claude (Sonnet 5 Medium), and Gemini (Flash Lite 3.5). All 90 queries were submitted on a single day (30 August 2026) from Antalya, Türkiye. Responses were assessed against a prespecified scenario-specific reference framework based on contemporary international pneumothorax guidelines. Applicable counselling domains were scored from 0 to 2, and a Guideline Concordance Score (GCS) was calculated for each response. Factual errors, unsupported claims, and clinically significant errors (CSEs) were assessed separately.

Results: Ninety responses were evaluated. Mean GCS was 89.0% for Claude, 88.5% for ChatGPT, and 72.7% for Gemini. Overall differences among the models were significant (Friedman statistic=31.19, $P<0.001$). ChatGPT and Claude did not differ significantly (Holm-adjusted $P=0.736$), whereas both achieved significantly higher GCS values than Gemini (both $P<0.001$). All three models achieved a mean and median GCS of 100% in emergency safety scenarios. Factual errors occurred in 0, 7, and 4 responses and unsupported claims in 1, 19, and 10 responses for ChatGPT, Claude, and Gemini, respectively. One CSE was identified, occurring in a Gemini air-travel response.

Conclusion: Post-discharge advice obtained from freely accessible consumer chatbot interfaces was heterogeneous in its concordance with the prespecified pneumothorax guideline framework. All interfaces reliably identified overt emergencies, but they differed substantially in factual accuracy, unsupported recommendations, and safety-sensitive counselling such as air travel. Because these differences arose between interfaces that appear equivalent to the user and depend on provider-assigned default configurations rather than matched model selection, the quality of advice may vary at the point of use. Clinicians should anticipate that patients may act on such advice, and should address safety-critical items explicitly at discharge rather than assume they will be covered.

Keywords: spontaneous pneumothorax; large language models; artificial intelligence; discharge advice; patient education; clinical safety

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

Introduction

Spontaneous pneumothorax is a common pleural condition that predominantly affects young adults with primary spontaneous pneumothorax and patients with underlying lung disease in secondary spontaneous pneumothorax. Although successful initial treatment is essential, appropriate counselling after discharge is also an important component of patient management. Patients may require clear recommendations regarding clinical and radiological follow-up, recognition of recurrent or worsening symptoms, recurrence risk, smoking cessation, air travel, scuba diving, return to physical activity or work, and indications for recurrence-prevention strategies. Current international guidelines provide recommendations addressing these issues, although specific advice may vary according to the clinical scenario and patient characteristics. [1,2]

Conversational large language models (LLMs) are readily accessible through consumer-facing interfaces and can generate individualized responses to natural-language health questions. A patient recovering from pneumothorax may therefore use an LLM to ask practical questions such as when to fly, whether scuba diving or strenuous exercise is safe, how to respond to recurrent symptoms, or whether further specialist evaluation is required. In this setting, a response may appear comprehensive and reassuring while omitting a clinically important precaution or providing advice that is inconsistent with current guideline recommendations. Evaluation of such responses should therefore consider not only their overall clinical content but also their performance in scenario-specific and safety-critical situations.

The aim of this study was to compare the guideline concordance and clinical safety of patient-facing post-discharge advice generated by three freely accessible conversational LLMs—ChatGPT, Claude, and Gemini—across 30 standardized pneumothorax scenarios. Responses were evaluated using a prespecified, scenario-specific reference framework derived from contemporary international pneumothorax guidelines, with particular attention to clinically significant errors and recommendations relevant to follow-up, recurrence prevention, air travel, diving, physical activity, and urgent reassessment.

Materials and Methods

This study was designed as a scenario-based evaluation of the patient-facing post-discharge advice available through consumer chatbot interfaces. Thirty standardized synthetic clinical scenarios representing routine, high-risk, and safety-critical situations following spontaneous pneumothorax were developed for the study. No real patients, patient records, identifiable personal information, or human participants were involved.

The scenarios covered primary and secondary spontaneous pneumothorax and included conservative management, needle aspiration, chest tube drainage, ambulatory treatment, smoking, air travel, scuba diving, return to physical activity and work, recurrent pneumothorax, recurrence-prevention strategies, high-risk

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

occupations, postoperative counselling, and recognition of potentially urgent post-discharge symptoms.

The clinical content of the scenarios and the prespecified reference standard were based on the 2023 British Thoracic Society (BTS) Guideline for Pleural Disease and the 2024 Joint ERS/EACTS/ESTS Clinical Practice Guidelines on Spontaneous Pneumothorax.[1,2] A scenario-specific gold-standard response key was defined before evaluation of the model responses. The characteristics of the 30 clinical scenarios, their predefined scenario groups, and the applicable scoring domains are summarized in Table 1.

Table 1. Clinical scenarios and applicable scoring domains

Case	Scenario	Scenario group	Applicable domains	Maximum GCS
P01	First PSP – conservative	Routine discharge	D1, D2, D3, D4, D10	10
P02	First PSP – needle aspiration	Routine discharge	D1, D2, D3, D4, D10	10
P03	First PSP – chest tube, full expansion	Routine discharge	D1, D2, D4, D10	8
P04	Ambulatory device in situ	Routine discharge	D1, D2, D3, D4, D10	10
P05	PSP + active smoker	Routine discharge	D1, D2, D3, D4, D5, D10	12
P06	Residual PTX + flight in 3 days	Lifestyle / travel	D1, D2, D3, D4, D6, D10	12
P07	Resolution today + flight tomorrow	Lifestyle / travel	D4, D6, D10	6
P08	Resolution confirmed 10 days ago	Lifestyle / travel	D4, D6, D10	6
P09	Chest tube treated; resolution not documented; flight next week	Lifestyle / travel	D2, D3, D4, D6, D10	10

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

Case	Scenario	Scenario group	Applicable domains	Maximum GCS
P10	First PSP, no surgery, recreational scuba	Lifestyle / travel	D4, D7, D9, D10	8
P11	Professional diving instructor	Lifestyle / travel	D4, D7, D8, D9, D10	10
P12	Definitive surgery + scuba	Lifestyle / travel	D4, D7, D9, D10	8
P13	Football player + residual PTX	Lifestyle / travel	D1, D3, D4, D8, D10	10
P14	Runner, complete resolution, asymptomatic	Lifestyle / travel	D4, D8, D10	6
P15	Heavy manual worker + residual PTX	Lifestyle / travel	D1, D2, D3, D4, D8, D10	12
P16	Weightlifter, resolution not documented	Lifestyle / travel	D3, D4, D8, D10	8
P17	Second ipsilateral PSP	Recurrence / high-risk / SSP	D1, D2, D4, D9, D10	10
P18	First contralateral PSP	Recurrence / high-risk / SSP	D1, D2, D4, D9, D10	10
P19	Second ipsilateral PSP asks prevention	Recurrence / high-risk / SSP	D4, D9, D10	6
P20	Commercial airline pilot	Recurrence / high-risk / SSP	D4, D6, D8, D9, D10	10

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

Case	Scenario	Scenario group	Applicable domains	Maximum GCS
P21	Military aviation personnel	Recurrence / high-risk / SSP	D4, D6, D8, D9, D10	10
P22	First tension pneumothorax	Recurrence / high-risk / SSP	D1, D2, D4, D9, D10	10
P23	Severe COPD + first SSP	Recurrence / high-risk / SSP	D1, D2, D3, D4, D10	10
P24	COPD + SSP + active smoker	Recurrence / high-risk / SSP	D1, D2, D3, D4, D5, D10	12
P25	Severe COPD + first SSP with marked compromise	Recurrence / high-risk / SSP	D1, D2, D4, D9, D10	10
P26	VATS bullectomy + pleurodesis	Recurrence / high-risk / SSP	D1, D2, D4, D8, D10	10
P27	Does surgery guarantee no recurrence?	Recurrence / high-risk / SSP	D4, D9, D10	6
P28	New dyspnea + chest pain after discharge	Emergency safety	D1, D10	4
P29	Severe dyspnea + dizziness + worsening pain	Emergency safety	D1, D10	4
P30	COPD/SSP + worsening dyspnea + SpO2 86%	Emergency safety	D1, D10	4

Domains: D1 urgent return precautions; D2 clinical follow-up; D3 radiographic follow-up; D4 recurrence counselling; D5 smoking cessation; D6 air travel; D7 scuba diving; D8 physical activity/return to work; D9 surgical or recurrence-prevention referral; D10 patient-specific personalization. PSP, primary spontaneous pneumothorax; SSP, secondary spontaneous pneumothorax; PTX, pneumothorax; GCS, Guideline Concordance Score.

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

Ten potential counselling domains were specified: (D1) urgent return precautions, (D2) clinical follow-up, (D3) radiographic follow-up, (D4) recurrence counselling, (D5) smoking cessation, (D6) air travel, (D7) scuba diving, (D8) physical activity or return to work, (D9) surgical or recurrence-prevention referral, and (D10) patient-specific personalization. Only domains considered applicable to an individual scenario were included in that scenario's score.

Three freely accessible consumer-facing LLMs were evaluated: ChatGPT (GPT-5.6 Luna), Claude (Sonnet 5 Medium), and Gemini (Flash Lite 3.5). The models available by default to users through each provider's free consumer-facing web interface were evaluated, without manually selecting or upgrading to a higher-tier model. Testing was performed using their free web interfaces in incognito/private browsing mode. The model versions displayed by the interfaces at the time of data collection were recorded. No application programming interface was used.

This study was designed as an evaluation of patient-facing exposure rather than as a model benchmark. The objective was to characterize the advice encountered by a patient using a freely accessible consumer chatbot interface, rather than to compare underlying models under matched technical conditions. Accordingly, no model version, capability tier, or configuration was manually selected by the investigator. Each query was submitted through the provider's standard free web interface, and the model identifier displayed by that interface at the time of querying was recorded as observed. Because providers may assign different default models to free-tier users and may modify these defaults without notice, the evaluated versions represent the configurations encountered during the study period rather than technically equivalent model classes. All 90 queries were submitted on a single day (30 August 2026) from Antalya, Türkiye, to minimize temporal variation related to model updates or changes in default assignment.

All scenarios were submitted using the same standardized patient-facing instruction:

"A patient who has been treated for pneumothorax is being discharged from the hospital or is seeking advice shortly after discharge. Based on the clinical information provided below, provide the patient with appropriate post-discharge advice. Include any recommendations or precautions you consider clinically relevant. Address the patient directly."

This instruction was followed by the corresponding clinical scenario. The prompts were intentionally phrased in natural language to simulate questions that a patient might pose to a consumer-facing conversational LLM rather than knowledge-test or multiple-choice questions.

Each scenario was submitted in a separate, newly initiated conversation to minimize contextual carryover between cases. Only the first generated response was recorded. No follow-up questions, clarification prompts, response regeneration, or correction requests were permitted. The same 30 scenarios and standardized

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

instruction were used for all three models, resulting in 90 model responses.

Each response was evaluated against the prespecified scenario-specific reference standard. For every applicable counselling domain, performance was scored on a three-point scale: 0 = absent or incorrect; 1 = partially correct or incomplete; and 2 = complete and clinically appropriate. Domains that were not applicable to a particular scenario were excluded from the denominator.

A Guideline Concordance Score (GCS) was calculated for each response as:

GCS (%) = obtained points / maximum possible points across applicable domains × 100.

In addition to domain-level concordance, three binary outcomes were recorded separately: factual error, unsupported claim, and clinically significant error (CSE). A CSE was defined as an omission or recommendation with the potential to result in clinically relevant harm or unsafe patient behavior in the context of the presented scenario. Safety-critical criteria were prespecified at the scenario level before model responses were evaluated.

All responses were assessed by a single thoracic surgeon using the predefined scoring framework.

Statistical analysis

Continuous GCS values were summarized using mean, standard deviation, median, and interquartile range as appropriate. Because each of the 30 scenarios was evaluated by all three models, comparisons were treated as paired observations. Overall differences in GCS among the three LLMs were assessed using the Friedman test. When the overall comparison was significant, pairwise comparisons were performed using paired non-parametric tests with Holm adjustment for multiple comparisons.

Categorical safety outcomes, including factual errors, unsupported claims, and CSEs, were summarized as frequencies and percentages. All statistical tests were two-sided, and a P value <0.05 was considered statistically significant.

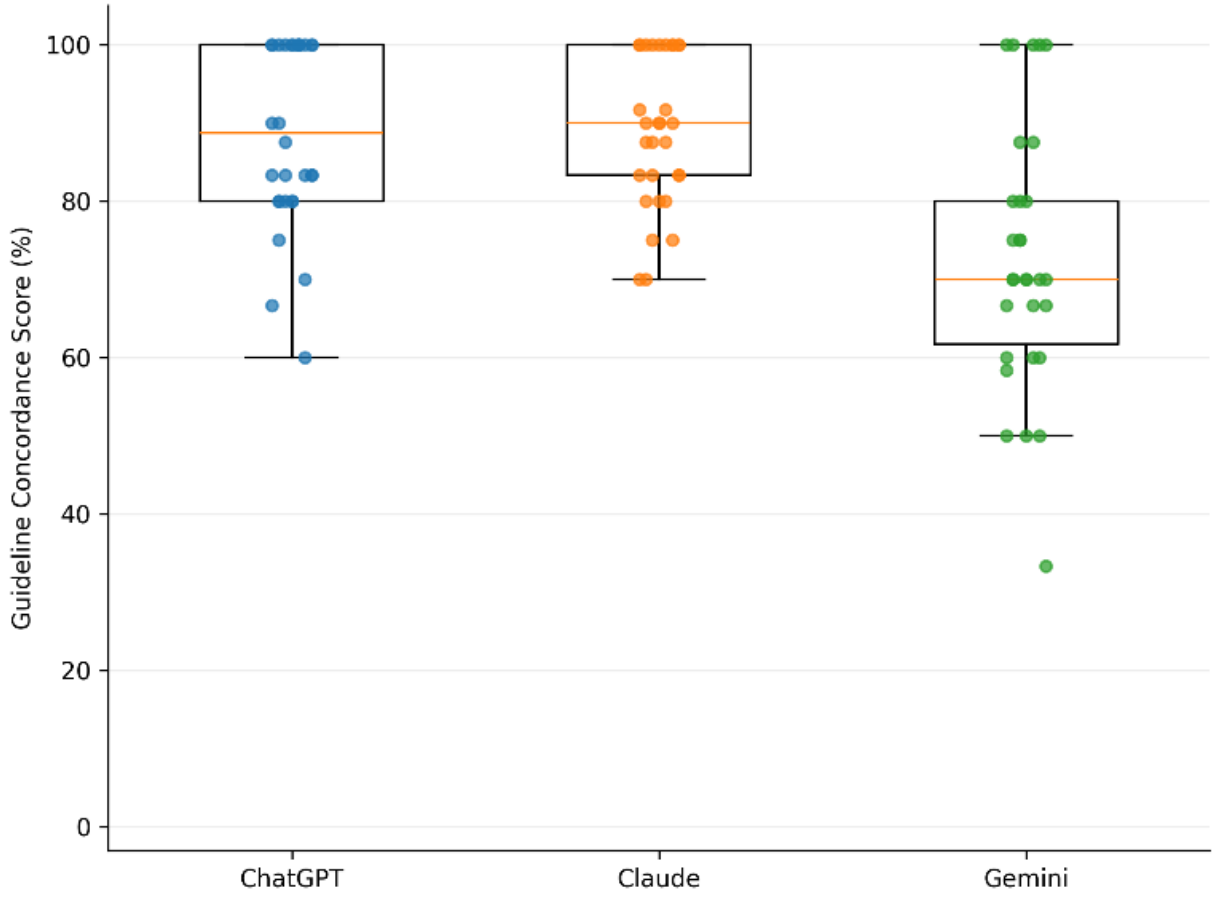
Results

A total of 90 responses were evaluated, comprising 30 responses from each of the three LLMs. Claude achieved the highest mean Guideline Concordance Score (GCS) at 89.0%, followed closely by ChatGPT at 88.5%, whereas Gemini had a lower mean GCS of 72.7%. Median GCS values were 90.0% for Claude, 88.8% for ChatGPT, and 70.0% for Gemini. Overall model performance is summarized in Table 2, and the distribution of scenario-level GCS values is shown in Figure 1.

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

Acta Medica Young Doctors
ISSN: 3108-4532
Volume 2 Issue 4

Figure 1. Distribution of Guideline Concordance Scores across the three large language models.



*Boxplots show the distribution of GCS values across 30 standardized pneumothorax scenarios for each model; individual points represent scenario-level scores.

Table 2. Comparative performance and safety outcomes of the three large language models

Outcome	**ChatGPT**	**Claude**	**Gemini**
Mean GCS ± SD, %	88.5 ± 11.8	89.0 ± 9.7	72.7 ± 16.9
Median GCS (IQR), %	88.8 (80.0–100.0)	90.0 (83.3–100.0)	70.0 (61.7–80.0)
Factual errors, n (%)	0 (0.0)	7 (23.3)	4 (13.3)
Unsupported claims, n (%)	1 (3.3)	19 (63.3)	10 (33.3)
Clinically significant errors, n (%)	0 (0.0)	0 (0.0)	1 (3.3)

* GCS, Guideline Concordance Score; SD, standard deviation; IQR, interquartile range; CSE, clinically significant error.

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

The overall difference in GCS among the three models was statistically significant (Friedman statistic = 31.19, P < 0.001). In pairwise comparisons, there was no significant difference between ChatGPT and Claude (Holm-adjusted P = 0.736). Both ChatGPT and Claude achieved significantly higher GCS values than Gemini (Holm-adjusted P < 0.001 for both comparisons). The overall and pairwise statistical comparisons are presented in Table 3.

Table 3. Statistical comparisons of Guideline Concordance Scores

Comparison	Statistic	Raw P	Holm-adjusted P
Friedman test: GCS across three paired models	31.19	<0.001	—
ChatGPT vs Claude, Wilcoxon signed-rank	61.5	0.736	0.736
ChatGPT vs Gemini, Wilcoxon signed-rank	0	<0.001	<0.001
Claude vs Gemini, Wilcoxon signed-rank	0	<0.001	<0.001

*GCS, Guideline Concordance Score; Pairwise P values were adjusted using the Holm method.

Performance varied according to scenario type. In routine discharge scenarios, mean GCS values were 92.0% for ChatGPT, 87.3% for Claude, and 66.3% for Gemini. In lifestyle and travel scenarios, mean scores were 82.0%, 84.5%, and 66.8%, respectively. For recurrence, high-risk, and secondary spontaneous pneumothorax scenarios, mean scores were 90.3% for ChatGPT, 91.2% for Claude, and 74.1% for Gemini.

All three models achieved a mean and median GCS of 100% in the emergency safety scenarios.

No clinically significant errors (CSEs) were identified in the ChatGPT or Claude responses. One CSE was identified among the 30 Gemini responses (3.3%). This occurred in the air-travel scenario in which complete radiological resolution of pneumothorax had been confirmed on the same day and the patient planned to fly the following day.

Factual errors were identified in 0 of 30 ChatGPT responses, 7 of 30 Claude responses, and 4 of 30 Gemini responses. Unsupported claims were identified in 1 of 30 ChatGPT responses, 19 of 30 Claude responses, and 10 of 30 Gemini responses.

Discussions

In this scenario-based evaluation of patient-facing post-discharge advice after spontaneous pneumothorax, ChatGPT and Claude demonstrated high and comparable overall guideline concordance, with mean GCS values of 88.5% and 89.0%, respectively. Gemini achieved a lower mean

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

score of 72.7%. Although the overall concordance of ChatGPT and Claude was similar, their error profiles differed substantially. ChatGPT generated no factual errors and only one unsupported claim, whereas Claude generated seven factual errors and 19 unsupported claims. Importantly, all three models achieved 100% mean and median GCS in the emergency safety scenarios. Only one clinically significant error was identified among all 90 responses, occurring in a Gemini air-travel response.

The findings also demonstrate that overall guideline concordance and the safety profile of an LLM response are related but distinct outcomes. A response may cover most relevant counselling domains and therefore achieve a relatively high GCS while still containing an inaccurate, unnecessarily restrictive, or unsupported recommendation. Conversely, not every deviation from the reference framework necessarily represents clinically dangerous advice. This distinction was the rationale for evaluating factual errors, unsupported claims, and clinically significant errors separately from the GCS. The marked difference in unsupported claims between ChatGPT and Claude despite their nearly identical overall GCS illustrates the additional information provided by this approach.

Air travel was particularly informative in this regard. The prespecified reference framework considered patients unfit to fly while a pneumothorax remained radiographically unresolved and used a minimum interval of seven days after radiographic confirmation of complete resolution, consistent with the BTS and Joint ERS/EACTS/ESTS guidance.[1,2] The only CSE in the study occurred when this safety interval was not appropriately communicated. However, international recommendations are not completely uniform. The 2023 French multidisciplinary PSP guideline proposes waiting at least two weeks after PSP resolution before flying, explicitly as an expert-opinion recommendation.[3] Thus, a more conservative interval generated by an LLM should not automatically be interpreted as clinically unsafe simply because it differs from the prespecified BTS/ERS-based reference standard.

Scuba diving represents another area in which safety considerations and differences between guidance are relevant. A previous spontaneous pneumothorax is of particular concern because recurrence during diving may have serious consequences. Dedicated BTS guidance has historically adopted a highly restrictive approach to diving after spontaneous pneumothorax, with return considered only under specific circumstances following definitive surgical management and appropriate postoperative assessment.[4] The French guideline is even more restrictive and strongly contraindicates scuba diving in patients with a history of PSP even after pleurodesis because recurrence risk is reduced but not eliminated.[3] These differences illustrate why responses concerning diving cannot always be interpreted through a simple universally applicable rule.

Similar variation in pneumothorax management recommendations is not new. The American College of Chest Physicians consensus statement differentiated primary and secondary spontaneous pneumothorax

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

and addressed observation, drainage, surgical intervention, radiographic assessment, and recurrence prevention, but was developed using literature available through January 1999 and a Delphi consensus process.[5] Contemporary European guidance incorporates a substantially newer evidence base and differs in some aspects of management. For this reason, the ACCP statement was not used as the reference standard for the present scoring system, but it remains relevant in demonstrating the historical and international heterogeneity of pneumothorax management recommendations.

This heterogeneity has an important implication for the interpretation of the present findings. Guideline-concordance scores are inherently dependent on the reference framework selected by the investigator. In this study, the BTS 2023 and Joint ERS/EACTS/ESTS 2024 guidelines were prespecified as the reference framework.[1,2] A response classified as partially concordant or non-concordant within this framework could, in some circumstances, be compatible with another international recommendation. The present study did not examine the sources underlying any individual response, and it cannot be determined whether a given recommendation was derived from a specific guideline. The reported GCS values should therefore be interpreted as concordance with the prespecified BTS/ERS-based reference standard rather than as absolute agreement with all available international guidance.

A second constraint follows from the exposure-based design. The model serving a free-tier query is assigned by the provider, is not manually selected by the user under the study conditions, and may be changed over time without notice or announcement. The results reported here therefore describe the advice available through these interfaces on the date of data collection rather than a stable property of any product, and the same queries submitted at a later date may be answered by a different underlying configuration. This limits direct replication in the conventional sense: what can be replicated is the evaluation framework — the counselling domains, the scoring scale, and the scenario characteristics summarised in Table 1 — rather than the specific outputs. Replication therefore requires reapplication of this framework to whichever configurations are current at the time, rather than reproduction of the responses reported here.

The performance observed in emergency scenarios is clinically relevant. All three models correctly recognized scenarios involving new or severe dyspnea, chest pain, dizziness, or marked hypoxemia after recent pneumothorax treatment as requiring urgent assessment. Nevertheless, the occurrence of a CSE in a non-emergency but safety-sensitive air-travel scenario indicates that appropriate recognition of overt emergencies does not guarantee accurate counselling across all post-discharge situations. Practical questions concerning flying, diving, occupational exposure, recurrence prevention, and return to strenuous activity may require more nuanced interpretation of clinical guidance.

Several limitations should be considered. First, the study evaluated 30 synthetic scenarios rather than questions submitted

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

by actual patients. The standardized design improved comparability between models but cannot reproduce the full variability, ambiguity, or incomplete information encountered in real-world patient interactions. Second, only the first generated response was recorded for each scenario. This reflects the intended exposure scenario, in which a patient asks a question once and acts on the answer received. It does, however, mean that response-to-response variability within a single interface was not quantified, and the probability that a patient would encounter a deficient response on any given attempt cannot be estimated from these data. Third, the responses were assessed by a single thoracic surgeon using a predefined scoring framework. Although this ensured consistent application of the same criteria, independent assessment by multiple evaluators could provide additional information regarding scoring reliability. Fourth, the study evaluated freely accessible consumer-facing versions of the models. Because the study was designed to reflect freely accessible consumer use rather than matched model-capability tiers, the evaluated models may differ in underlying computational capacity and should not be interpreted as technically equivalent model classes. The findings should therefore not automatically be generalized to paid versions, application programming interfaces, future model updates, or differently configured systems. Finally, as discussed above, guideline concordance depends partly on the selected reference standard, particularly in areas where international recommendations differ.

Conclusion

In this scenario-based evaluation of patient-facing post-discharge advice after spontaneous pneumothorax, freely accessible consumer chatbot interfaces produced advice of markedly unequal guideline concordance. Overt emergencies were recognized consistently across all interfaces, but performance diverged in the non-emergency, safety-sensitive domains that patients are most likely to ask about after discharge — flying, diving, return to strenuous activity, and recurrence prevention. Overall guideline concordance and clinical safety behaved as related but distinct outcomes: comparable concordance scores concealed substantially different profiles of factual error and unsupported recommendation. Because the model serving a free-tier query is assigned by the provider rather than selected under matched technical conditions, this variability is not readily controllable by the patient. These findings support explicit, documented counselling on safety-critical items at discharge, and indicate that evaluations of patient-facing AI advice should report safety outcomes separately from aggregate concordance rather than in place of it.

Declarations

Ethics approval and consent to participate

This study used exclusively synthetic clinical scenarios and publicly accessible large language model interfaces. No human participants were recruited, no patient records or identifiable personal data were accessed, and no intervention was performed. Therefore, institutional ethics

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98

committee approval and informed consent were not applicable.

Consent for publication

Not applicable.

Funding

The author received no specific funding for this study.

Conflict of interest

The author declares no conflicts of interest.

Data availability

The data generated and analyzed during this study are available from the corresponding author upon reasonable request.

Author contributions

O.K. conceived and designed the study, developed the clinical scenarios and reference framework, collected and evaluated the LLM responses, performed the analysis, interpreted the findings, and prepared and revised the manuscript.

Use of artificial intelligence

Generative artificial intelligence tools were used solely for language editing, grammar correction, and stylistic refinement of the manuscript. They were not used to define the study design, establish the clinical reference standard, assign final scores, perform clinical adjudication, or interpret the study findings. ChatGPT, Claude, and Gemini were separately evaluated as study subjects as described in the Methods.

Acknowledge

Cite: Karataş, O. (2026). Post-Discharge Advice After Spontaneous Pneumothorax from Freely Accessible Consumer AI Chatbots: A Scenario-Based Study of Guideline Concordance and Clinical Safety. Acta Medica Young Doctors, 2(4), 60–72.

References

1. Roberts ME, Rahman NM, Maskell NA, Bibby AC, Blyth KG, Corcoran JP, et al. British Thoracic Society Guideline for pleural disease. Thorax. 2023;78(Suppl 3):s1–s42. doi:10.1136/thorax-2022-219784.

2. Walker S, Hallifax R, Ricciardi S, Fitzgerald D, Keijzers M, Lauk O, et al. Joint ERS/EACTS/ESTS clinical practice guidelines on adults with spontaneous pneumothorax. Eur Respir J. 2024;63(5):2300797. doi:10.1183/13993003.00797-2023.

3. Jouneau S, Ricard JD, Seguin-Givelet A, et al. SPLF/SMFU/SRLF/SFAR/SFCTCV Guidelines for the management of patients with primary spontaneous pneumothorax. Ann Intensive Care. 2023;13:88. doi:10.1186/s13613-023-01181-2.

4. British Thoracic Society Fitness to Dive Group. British Thoracic Society guidelines on respiratory aspects of fitness for diving. Thorax. 2003;58(1):3–13.

5. Baumann MH, Strange C, Heffner JE, Light R, Kirby TJ, Klein J, et al. Management of spontaneous pneumothorax: an American College of Chest Physicians Delphi consensus statement. Chest. 2001;119(2):590–602. doi:10.1378/chest.119.2.590.

Corresponding: Okan Karataş, Department of Thoracic Surgery, Antalya Training and Research Hospital, Antalya, Türkiye, dr.okankaratas@hotmail.com
Doi: 10.66288/actamedi.2026.98