УДК 616.97:004.8

DOI: http://doi.org/10.30978/TB2025-4-114

A.Ye. Bogomolov[1], O.B. Bondarchuk[1], L.M. Kyrychenko[1], I.V. Korytska[1], S.V. Zaikov[2]

[1] National Pirogov Memorial Medical University, Vinnytsya, Ukraine
[2] Shupyk National Healthcare University of Ukraine, Kyiv, Ukraine

Artificial Intelligence-Based Chatbots May Influence Decision-Making in Patients with Respiratory and Non-Respiratory Allergic Diseases

Objective — to assess the potential influence of artificial intelligence-generated responses on decision-making and care pathways in patients with respiratory and non-respiratory allergic conditions.

Materials and methods. Twelve questions were submitted to two of the most widely used artificial intelligence-based chatbots: ChatGPT-4o and Gemini 2.0 Flash. Half of these questions were developed through an analysis of Google Trends data from the past year in Ukraine (October 2023—September 2024). Another half were compiled from an online survey of practising physicians, who identified the most frequently asked questions by patients during clinical consultations. Five experts independently assessed each chatbot's responses based on three parameters: accuracy, correctness, and comprehensiveness, using a 0—3 scale (0 = completely inaccurate/incorrect/non-comprehensive, 3 = fully accurate/correct/comprehensive). Later, all these topics were categorised into two blocks: Decision-Making block and Awareness block.

Results and discussion. The mean scores were as follows: ChatGPT achieved 2.08 ± 0.46 (accuracy), 2.07 ± 0.52 (correctness) and 2.10 ± 0.57 (comprehensiveness) points, while Gemini scored 1.97 ± 0.71 (accuracy), 2.00 ± 0.69 (correctness) and 2.05 ± 0.67 (comprehensiveness) points. These results indicate a slight overall advantage for ChatGPT, with the largest difference observed in accuracy (0.11 points). Statistical analysis indicated moderate to strong agreement between experts, which is generally sufficient to validate the results.

The analysis revealed that Decision-Making block questions were answered by ChatGPT with accuracy = 2.20 ± 0.52, correctness = 2.17 ± 0.50, comprehensiveness = 2.26 ± 0.48 points, and by Gemini with accuracy = 2.00 ± 0.58, correctness = 1.91 ± 0.60, comprehensiveness = 2.06 ± 0.55 points. Awareness block questions were answered by ChatGPT with accuracy = 1.92 ± 0.60, correctness = 1.92 ± 0.58 and comprehensiveness = = 1.88 ± 0.62 points, and by Gemini with accuracy = 1.88 ± 0.65, correctness = 2.12 ± 0.57 and comprehensiveness = 2.04 ± 0.59 points.

Conclusions. The experts in our study evaluated the answers on a four-point scale (from 0 to 3 points), and both chatbots answered on average 2 points out of 3 possible for all parameters — accuracy, correctness and completeness, which is a reasonable indicator. However, the analysis clearly showed (and this is noticeable in the average deviation) that the range of Gemini's answer scores was higher than ChatGPT's, that is, the chatbot gave both more high-quality and low-quality answers. This increases the chance for the questioner to receive both a bad and a good answer, which is an indicator of the chatbot's lower predictability.

In the Decision-Making block, ChatGPT was statistically significantly better, but in the Awareness block, ChatGPT was better only in the accuracy of answers, while Gemini statistically significantly answered questions more completely and correctly according to expert assessments. ChatGPT consistently outperformed Gemini in the Decision-Making block, indicating its suitability for tasks requiring structured decision-making. In contrast, Gemini outperformed in the Awareness block, especially in correctness and completeness, indicating its effectiveness for information queries.

Keywords

Respiratory and non-respiratory allergic diseases, artificial intelligence, chatbot.

Since the public release of the ChatGPT language model nearly 2.5 years ago, there has been an unprecedented pace of technology development and engagement and integration of artificial intelligence (AI) technologies into everyday life. Being the fastest-growing consumer software application in history having gained over 100 million users in just the first two months by January 2023, ChatGPT has reached over 400 million weekly active users globally in February 2025 [4, 5].

One of the most prominent shifts has been the widespread incorporation of conversational AI systems — such as virtual assistants and chatbots — into websites, applications and online services. These systems are increasingly used to streamline user interactions, answer questions and provide recommendations. Medical websites are no exception to this trend. Many now offer AI-powered support tools to help users interpret symptoms, suggest possible diagnoses or direct them to the appropriate healthcare professionals.

As a result, when individuals seek medical advice or try to understand their condition — for example, by asking which specialist to consult or how to manage certain symptoms — they are increasingly likely to receive a response generated by an AI chatbot rather than a human expert. This growing reliance on AI for preliminary medical information raises important concerns about the accuracy, completeness and clinical reliability of such responses, particularly in domains like allergy care, where misinformation can have significant consequences.

Objective — to assess the potential influence of AI-generated responses on decision-making and care pathways in patients with respiratory and non-respiratory allergic conditions.

Specifically, we evaluated the quality of responses provided by two of the most popular publicly available AI models (as of October 2024), ChatGPT-4o and Gemini 2.0 Flash, to determine how effectively they address common allergy-related questions and to explore the implications of their use in patient-facing digital environments.

Materials and methods

Twelve questions were submitted to two of the most widely used AI-based chatbots: ChatGPT-4o and Gemini 2.0 Flash. Half of these questions were developed through an analysis of Google Trends data from the past year in Ukraine (October 2023—September 2024), focusing on three key topics — «allergies», «allergens» and «allergy tests». The corresponding data with popularity trends are shown in Fig. 1.

The remaining six questions were compiled from an online survey of practising physicians, who identified the most frequently asked questions by patients during clinical consultations. Later all these topics were categorised into two blocks: Decision-Making block (questions 2, 5, 6, 9, 10, 11, 12) and Awareness block (questions 1, 3, 4, 7, 8). Five experts independently assessed each chatbot's responses based on three parameters: accuracy, correctness, and comprehensiveness, using a 0—3 scale (0 = completely inaccurate/incorrect/non-comprehensive 3 = fully accurate/correct/comprehensive). Each answer was evaluated according to three parameters:

- Accuracy — a measure of how accurately the answer to the formulated request/question is given — is there «water», unnecessary information etc.;
- Correctness — a measure of whether the information provided, in the opinion of the expert, is correct;
- Comprehensiveness — a measure of how comprehensive the answer to the question is, whether something important needs to be added, what is missing etc.

This resulted in a total of 60 evaluations per parameter (5 experts — 12 questions). Each of the five experts was a practising allergist (with a minimum of 10 years of clinical experience), ensuring expertise in understanding responses to questions and a university faculty member (with a minimum of 15 years of teaching experience), ensuring expertise in assessing responses from an educational perspective.

Descriptive statistics, including mean scores, were calculated for each parameter and block. The Mann—Whitney U test was employed to compare the performance of ChatGPT and Gemini, with a significance level of $p < 0.05$. Inter-rater agreement was initially assessed using Kendall's W-coefficient, though subsequent analysis utilised the Intraclass Correlation Coefficient (ICC) due to the numerical nature of the ratings. Visualisations, including bar charts, radar charts and bubble charts (heatmaps), were generated to illustrate the results.

Results and discussion

The mean scores across all 12 questions were calculated as follows: ChatGPT achieved 2.08 ± 0.46 (accuracy), 2.07 ± 0.52 (correctness) and 2.10 ± ± 0.57 (comprehensiveness) points, while Gemini scored 1.97 ± 0.71 (accuracy), 2.00 ± 0.69 (correctness) and 2.05 ± 0.67 (comprehensiveness) points. These results indicate a slight overall advantage for ChatGPT, with the largest difference observed in accuracy (0.11 points) (Fig. 2).

The analysis shows that the intervals for ChatGPT are more consistent and have higher lower bounds,

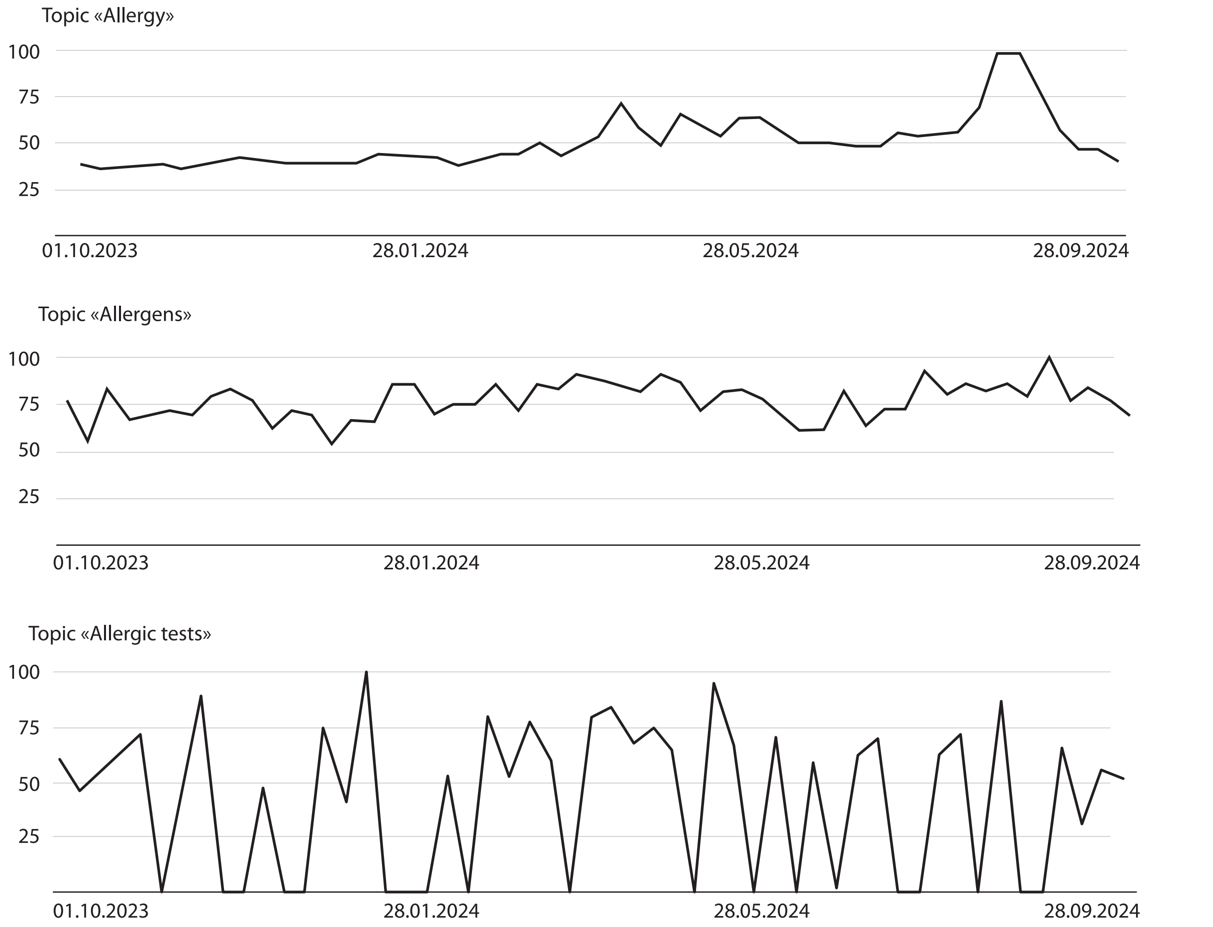

Fig. 1. **Google trends on requests from Oct 2023—Sep 2024**

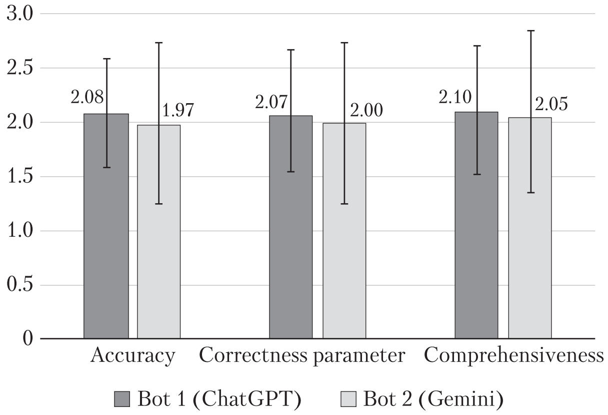

Fig. 2. **Average scores of the chatbots with standard deviation**

Table. **Results of the calculation of Kendall's W coefficient for assessing expert agreement**

Parameter	Model	Kendall's W
Accuracy	ChatGPT	0.6627
	Gemini	0.6901
Correctness	ChatGPT	0.5368
	Gemini	0.6386
Comprehensiveness	ChatGPT	0.4920
	Gemini	0.6151

while the Gemini responses were estimated with wider intervals and lower mean values, indicating greater variability in expert assessments.

The Mann—Whitney U test was conducted to compare the performance of ChatGPT and Gemini across all parameters. Significant differences were observed: accuracy ($p \approx 0.03$, $p < 0.05$), correctness ($p \approx 0.04$, $p < 0.05$) and comprehensiveness ($p \approx 0.02$, $p < 0.05$). These results indicate that ChatGPT outperforms Gemini across all parameters.

To check the consistency of the experts' assessment of the answers to the questions, we calculated the Kendall's W coefficient. The results are shown in Table.

As can be seen from Table 1, Kendall's W values between 0.5 and 0.7 indicate moderate to strong agreement between experts, which is generally sufficient to validate the results. The highest agreement was observed for accuracy, especially with the Gemini chatbot.

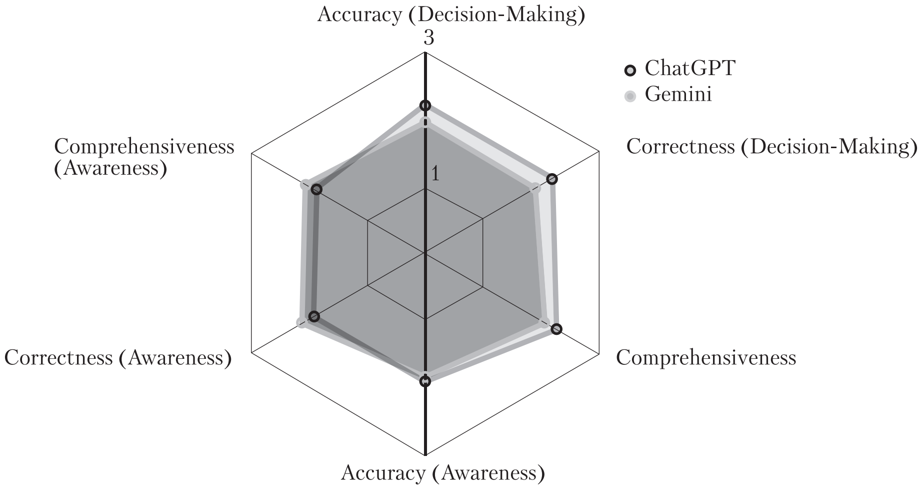

Fig. 3. **Comparison of chatbots' response scores by question blocks**

As already indicated in the research methodology, for the analysis of questions we classified them as a Decision-Making block (questions the answers to which can potentially affect the patient's decision-making) and an Awareness block (questions the answers to which can potentially affect the patient's awareness) (Fig. 3). The analysis revealed that Decision-Making block questions were answered by ChatGPT with accuracy = 2.20 ± 0.52, correctness = 2.17 ± 0.50, comprehensiveness = = 2.26 ± 0.48 points, and by Gemini with accuracy = 2.00 ± 0.58, correctness = 1.91 ± 0.60, comprehensiveness = 2.06 ± 0.55 points. The Mann—Whitney test revealed significant differences in accuracy ($p \approx 0.02$), correctness ($p \approx 0.01$) and comprehensiveness ($p \approx 0.03$), all with $p < 0.05$, confirming the superiority of ChatGPT in this block. Awareness block questions were answered by ChatGPT with accuracy = 1.92 ± 0.60, correctness = 1.92 ± 0.58, comprehensiveness = 1.88 ± 0.62 points, and by Gemini with accuracy = 1.88 ± 0.65, correctness = 2.12 ± 0.57, comprehensiveness = 2.04 ± 0.59 points. The Mann—Whitney test revealed non-significant differences for accuracy ($p \approx 0.07$, $p > 0.05$), but significant differences for correctness ($p \approx 0.04$) and comprehensiveness ($p \approx 0.03$), both with $p < 0.05$, where Gemini outperforms ChatGPT.

The radar chart shows ChatGPT's larger area in the Decision-Making block, indicating its superiority, while Gemini's larger area in the Awareness block (e.g., correctness 2.12) reflects its strengths.

Artificial intelligence language models are becoming increasingly popular, actively integrating into various areas of human life. Medicine is no exception, and there is a lot of research on the use of AI in student education [1, 3], medical programming [6] and even contributing to clinical decision-making [2]. All of these studies look at AI from the perspective of medical professionals, and the focus is primarily on improving the training of specialists or the provision of medical care. However, we were interested in the application of chatbots from the patient's perspective, because an increasing number of search engines and Internet services integrate their own or well-known language models to answer questions (for example, the Google search service already displays the results its own language model, Gemini, in response to a search query above the search results).

Conclusions

The experts in our study evaluated the answers on a four-point scale (from 0 to 3 points), and both chatbots answered on average 2 points out of 3 possible for all parameters — accuracy, correctness and completeness, which is a good indicator. This suggests that, in general, such answers were well formulated and argued. However, the analysis clearly showed (and this is noticeable in the average deviation) that the range of Gemini's answer scores was higher than ChatGPT's, that is, the chatbot produced both more high-quality and low-quality answers. This increases the likelihood that the questioner may receive either a poor or a good answer, indicating the chatbot's lower predictability.

The results of the analysis of answers by blocks proved to be more interesting. In the Decision-Making block, ChatGPT was answered questions significantly more completely and correctly according to expert assessments. ChatGPT consistently outperforms Gemini in the Decision-Making block, indicating its suitability for tasks requiring structured decision-making. In contrast, Gemini outperforms in the Awareness block, especially in correctness and completeness, indicating its effectiveness for information queries.

The authors acknowledge the limitations of the study. Chatbots are constantly improving, learning from machine learning and billions of user queries, and new versions of the language models we used are likely to perform better. Competing products are also actively developing, and their number is increasing daily. Expert ratings also contain an element of subjectivity, although the rating was blind (the experts did not know which chatbots they were evaluation) and statistical analysis showed a moderate level of agreement between the ratings.

Generative AI models hold promise for widespread use. However, our research has shown that their answers to user questions are still far from ideal, and the lack of predictive quality in the response can potentially influence both patient decision-making and awareness of various issues.

There is no conflict of interest.
Participation of authors: concept and design of the study — A.Ye. Bogomolov; collection of material — O.B. Bondarchuk, I.V. Korytska, L.M. Kyrychenko, S.V. Zaikov; processing of material, statistical data, writing of the text — A.Ye. Bogomolov; editing of the text — S.V. Zaikov, A.Ye. Bogomolov.

References

1. Burisch C, Bellary A, Breuckmann F, et al. ChatGPT-4 Performance on german continuing medical education — friend or foe (trick or treat)? Protocol for a randomized controlled trial. JMIR Res Protoc. 2025 Feb 6;14:e63887. doi: 10.2196/63887. PMID: 39913914; PMCID: PMC11843049.
2. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023 Jul 3;330(1):78-80. doi: 10.1001/jama.2023.8288. PMID: 37318797; PMCID: PMC10273128.
3. Kim TW. Application of artificial intelligence chatbots, including ChatGPT, in education, scholarly work, programming, and content generation and its prospects: a narrative review. J Educ Eval Health Prof. 2023;20:38. doi: 10.3352/jeehp.2023.20.38. Epub 2023 Dec 27. PMID: 38148495; PMCID: PMC11893184.
4. Liang C. ChatGPT sets record for fastest-growing user base - analyst note. Reuters [Internet]. 2023 Feb 1 [cited 2024 Jul 3]. Available from: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.
5. Liang C. OpenAI's weekly active users surpass 400 million in 2025. Reuters [Internet]. 2025 Feb 20 [cited 2024 Jul 3]. Available from: https://www.reuters.com/technology/artificial-intelligence/openais-weekly-active-users-surpass-400-million-2025-02-20/.
6. Tam W, Huynh T, Tang A, Luong S, Khatri Y, Zhou W. Nursing education in the age of artificial intelligence powered Chatbots (AI-Chatbots): Are we ready yet? Nurse Educ Today. 2023 Oct;129:105917. doi: 10.1016/j.nedt.2023.105917. Epub 2023 Jul 18. PMID: 37506622.

А.Є. Богомолов[1], О.Б. Бондарчук[1], Л.М. Кириченко[1], І.В. Корицька[1], С.В. Зайков[2]

[1] Вінницький національний медичний університет імені М.І. Пирогова
[2] Національний університет охорони здоров'я України імені П.Л. Шупика, Київ

Чат-боти на основі штучного інтелекту можуть впливати на прийняття рішень у пацієнтів із респіраторними та нереспіраторними алергійними захворюваннями

Мета роботи — оцінити потенційний вплив відповідей, згенерованих штучним інтелектом, на прийняття рішень і шляхи медичного догляду в пацієнтів із респіраторними та нереспіраторними алергійними станами.

Матеріали та методи. Дванадцять запитань було надано двом із найпоширеніших чат-ботів на основі штучного інтелекту: ChatGPT-4o та Gemini 2.0 Flash. Половина цих запитань була розроблена на основі аналізу даних Google Trends за останній рік в Україні (жовтень 2023 р. — вересень 2024 р.), інша половина була складена на основі онлайн-опитування практикуючих лікарів, які визначили найчастіші запитання, які ставлять пацієнти під час клінічних консультацій. П'ять експертів незалежно оцінювали відповіді кожного чат-бота за трьома параметрами: точність, правильність і вичерпність, використовуючи шкалу від 0 до 3 (0 = повністю неточно/неправильно/невичерпно, 3 = повністю точно/правильно/вичерпно). Пізніше всі теми були розподілені на два блоки: Прийняття рішень і Обізнаності.

Результати та обговорення. Середні оцінки були такими: ChatGPT отримав (2,08 ± 0,46) бала (точність), (2,07 ± 0,52) бала (правильність) і (2,10 ± 0,57) бала (вичерпність), тоді як Gemini — відповідно (1,97 ± 0,71), (2,00 ± 0,69) та (2,05 ± 0,67) бала. Ці результати вказують на незначну загальну перевагу ChatGPT із найбільшою різницею за точністю (0,11). Статистичний аналіз виявив помірну до високої узгодженість між експертами, що загалом є достатнім для підтвердження результатів.

Аналіз показав, що запитання блоку Прийняття рішень були оцінені для ChatGPT з такими показниками: точність — (2,20 ± 0,52) бала, правильність — (2,17 ± 0,50) бала, вичерпність — (2,26 ± 0,48) бала, для Gemini: точність — (2,00 ± 0,58) бала, правильність — (1,91 ± 0,60) бала, вичерпність — (2,06 ± 0,55) бала. Запитання блоку Обізнаності отримали оцінки для ChatGPT: точність — (1,92 ± 0,60) бала, правильність — (1,92 ± 0,58) бала, вичерпність — (1,88 ± 0,62) бала, для Gemini: точність — (1,88 ± 0,65) бала, правильність — (2,12 ± 0,57) бала, вичерпність — (2,04 ± 0,59) бала.

Висновки. Експерти в нашому дослідженні оцінювали відповіді за 4-бальною шкалою (від 0 до 3 балів). Обидва чат-боти в середньому отримали 2 бали з 3 можливих за всіма параметрами (точність, правильність і вичерпність), що є добрим показником. Проте аналіз чітко показав (що помітно за середнім відхиленням), що діапазон оцінок відповідей Gemini був вищим, ніж у ChatGPT, тобто цей чат-бот давав як більш якісні, так і менш якісні відповіді. Це підвищує ймовірність для того, хто запитує, отримати як погану, так і хорошу відповідь, що є показником гіршої передбачуваності чат-бота.

У блоці Прийняття рішень ChatGPT виявився статистично значущо кращим, але в блоці Обізнаності ChatGPT перевершував лише за точністю відповідей, тоді як, за оцінками експертів, Gemini статистично значущо відповідав повніше та правильніше. ChatGPT стабільно перевершував Gemini в блоці Прийняття рішень, що свідчить про його придатність для завдань, які потребують структурованого прийняття рішень. Натомість Gemini перевершував у блоці Обізнаності, особливо за правильністю та вичерпністю, що вказує на його ефективність для інформаційних запитів.

Ключові слова: респіраторні та нереспіраторні алергійні захворювання, штучний інтелект, чат-бот.

Контактна інформація / *Corresponding author*

Богомолов Артемій Євгенійович, д. мед. н., проф. кафедри фтизіатрії з курсом клінічної імунології та алергології
https://orcid.org/0000-0002-5336-4858
E-mail: bogomolov@vnmu.edu.ua

Стаття надійшла до редакції/*Received* 02.06.2024.
Стаття рекомендована до опублікування/*Accepted* 23.07.2025.
Стаття опублікована / *Published* 30.10.2025.

ДЛЯ ЦИТУВАННЯ

- Bogomolov AYe, Bondarchuk OB, Kyrychenko LM, Korytska IV, Zaikov SV. Artificial Intelligence-Based Chatbots May Influence Decision-Making in Patients with Respiratory and Non-Respiratory Allergic Diseases. Туберкульоз, легеневі хвороби, ВІЛ-інфекція. 2025;4:114-119. doi: 10.30978/TB2025-4-114.
- Bogomolov AYe, Bondarchuk OB, Kyrychenko LM, Korytska IV, Zaikov SV. Artificial Intelligence-Based Chatbots May Influence Decision-Making in Patients with Respiratory and Non-Respiratory Allergic Diseases. Tuberculosis, Lung Diseases, HIV Infection (Ukraine). 2025;4:114-119. http://doi.org/10.30978/TB2025-4-114.