European Journal of Therapeutics
pISSN: 2564-7784
eISSN: 2564-7040

Eur J Ther. 2024;30(6):900-909.
https://doi.org/10.58600/eurjther2569

Original Research

Evaluation of the Readability, Understandability, and Accuracy of Artificial Intelligence Chatbots in Terms of Biostatistics Literacy

İlkay Doğan[1,*], Pınar Günel[2], İhsan Berk[2], Buket İpek Berk[3]

[1] Department of Biostatistics, Faculty of Medicine, Gaziantep University, Gaziantep, Türkiye
[2] Department of Biostatistics, Faculty of Medicine, SANKO University, Gaziantep, Türkiye
[3] Department of Biostatistics, Graduate Education Institute, SANKO University, Gaziantep, Türkiye

Received: 2024-12-02 **Accepted:** 2024-12-23 **Published Online:** 2024-12-30

Corresponding Author
İlkay Doğan, Assoc. Prof., PhD
Address: Gaziantep University, Faculty of Medicine, Department of Biostatistics, Gaziantep, Türkiye
E-mail: ilkay_dgn58@hotmail.com

This study was presented as an oral presentation at the 25th National and 8th International Biostatistics Congress held in Sakarya on 18-20 November 2024.

© 2024, European Journal of Therapeutics, Gaziantep University School of Medicine.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ABSTRACT

Objective: Chatbots have been frequently used in many different areas in recent years, such as diagnosis and imaging, treatment, patient follow-up and support, health promotion, customer service, sales, marketing, information and technical support. The aim of this study is to evaluate the readability, understandability, and accuracy of queries made by researchers in the field of health through artificial intelligence chatbots in biostatistics.

Methods: A total of 10 questions from the topics frequently asked by researchers in the field of health in basic biostatistics were determined by 4 experts. The determined questions were addressed to the artificial intelligence chatbots by one of the experts and the answers were recorded. In this study, free versions of most widely preferred ChatGPT4, Gemini and Copilot chatbots were used. The recorded answers were independently evaluated as "Correct", "Partially correct" and "Wrong" by three experts who blinded to which chatbot the answers belonged to. Then, these experts came together and examined the answers together and made the final evaluation by reaching a consensus on the levels of accuracy. The readability and understandability of the answers were evaluated with the Ateşman readability formula, Sönmez formula, Çetinkaya-Uzun readability formula and Bezirci-Yılmaz readability formulas.

Results: According to the answers given to the questions addressed to the artificial intelligence chatbots, it was determined that the answers were at the "difficult" level according to the Ateşman readability formula, "insufficient reading level" according to the Çetinkaya-Uzun readability formula, and "academic level" according to the Bezirci-Yılmaz readability formula. On the other hand, the Sönmez formula gave the result of "the text is understandable" for all chatbots. It was determined that there was no statistically significant difference (p=0.819) in terms of accuracy rates of the answers given by the artificial intelligence chatbots to the questions.

Conclusion: It was determined that although the chatbots tended to provide accurate information, the answers given were not readable, understandable and their accuracy levels were not high.

Keywords: Artificial Intelligence, Chatbots, Biostatistics Literacy, Readability, Understandability, Accuracy.

INTRODUCTION

Artificial intelligence (AI), which we have heard about frequently in recent years, is renewing itself day by day. Artificial intelligence can be defined as a technology of a computer or machine that has features such as learning, problem solving, decision making, and sometimes imitating human intelligence and thoughts. The basis of the concept of artificial intelligence is based on intelligent machines that first appeared with the question "Can machines think?" by Turing [1]. The term artificial intelligence was first used by McCarthy et al. [2] in 1955. Samuel [3] programmed the computer using two machine learning procedures, which provides learning how to play checkers. In addition to these, the historical development process of artificial intelligence can be examined in detail in the studies of Pirim [4], and Öztürk and Şahin [5]. Two events that can be considered as milestone for the concept and applications of artificial intelligence: artificial intelligence beating world chess champion Garry Kasparov in 1997, and in 2015, AlphaGo, an artificial intelligence developed by Google, beating a professional Go player without giving them an advantage. These successes of artificial intelligence have shown the potential of artificial neural networks and deep learning, and have enabled machine learning algorithms to find a place in everyday life [6]. Today, artificial intelligence is encountered in almost every field of science. It plays an active role in the early diagnosis and treatment of diseases, especially in the fields of medicine and health (radiology [7,8], oncology [9,10], cardiology [11,12], gastroenterology [13,14], ophthalmology [15,16], surgery [17,18], etc.).

Chatbots, a sub-branch of artificial intelligence, are products of the field of natural language processing (NLP). Chatbots are trained with various databases to answer questions posed to them by users [19]. Chatbots can understand, interpret, and answer users' questions via text or voice to simulate human-like conversation using natural language processing (NLP) [20]. In addition, chatbots can answer consecutive questions, accept errors in their answers and correct themselves with reinforcement learning, understand and answer different languages, and refuse to answer inappropriate questions [19]. Chatbots are constantly improving themselves with the machine learning algorithms in their background to provide correct answers and perform better. The most preferred chatbots by users are ChatGPT (OpenAI), Copilot (Microsoft), Gemini (Google) due to their free versions and easy accessibility.

Main Points

- Although chatbots tend to provide accurate information, the results showed they were not readable, understandable, and had low accuracy levels.
- One of the most important limitations of chatbots is the lack of evidence-based information sources. Therefore, it is very important to check their answers.

Chatbots have been frequently used in many different areas in recent years, such as diagnosis and imaging, treatment, patient follow-up and support, health promotion, customer service, sales, marketing, information and technical support [21-26]. Despite this, there are still many question marks on the responses given by chatbots, and researchers are evaluating the readability, understandability and accuracy of the responses given by chatbots [20, 27-32]. In this study, the responses given by chatbots are evaluated for the first time in terms of biostatistical literacy. The aim of this study is to evaluate the readability, understandability, and accuracy of queries made by researchers in the field of health through artificial intelligence chatbots in biostatistics. In this context, it is thought that the results obtained will be guiding in the use of chatbots, especially in the field of health sciences.

MATERIALS AND METHODS

A total of 10 questions from topics that are thought to be frequently questioned by researchers in the field of health in basic biostatistics (deciding on appropriate statistical tests, interpretation of the results of applied tests, some basic statistical definitions, sample size calculation-power analysis, etc.) were determined by 4 experts (Table 1). The determined questions were addressed to artificial intelligence chatbots by one of the experts and the answers were recorded. In this study, free versions of most widely preferred ChatGPT4, Gemini and Copilot chatbots were used. The recorded answers were independently evaluated as "Correct", "Partially correct" and "Wrong" by three experts who blinded to which chatbot the answers belonged to. Then, these experts came together and examined the answers together and made the final evaluation by reaching a consensus on the accuracy of the answers. In addition, the readability and understandability of the answers were assessed by using the Ateşman readability formula,

Sönmez formula, Çetinkaya-Uzun readability formula and Bezirci-Yılmaz readability formulas.

Ateşman Readability Formula

The Ateşman readability formula was proposed by Ateşman in 1997 to assess the readability of Turkish texts. This measure is an adaptation of the formula proposed by Flesch in 1948. Emphasizing the difference in average word and sentence lengths between English and Turkish, Ateşman adapted the variables of the Flesch formula to Turkish sentence and word lengths and created his own formula [33].

$$\text{Readability score} = 198.825 - 40.175(x1) - 2.610(x2)$$

$$x1 = \frac{\text{(Syllable count)}}{\text{(Word count)}}$$

$$x2 = \frac{\text{(Word count)}}{\text{(Number of sentences)}}$$

According to the Ateşman readability formula, readability levels are scored as "1-29: very difficult", "30-49: difficult", "50-69: medium difficulty", "70-89: easy", "90-100: very easy" [33].

Sönmez Formula

In his study, Sönmez (2003) found that the Fog index used to evaluate the understandability of texts gave invalid results on Turkish texts and stated that the prerequisite for the understandability of a text is to know the meaning of the words. In this context, he created a formula based on unknown words for Turkish [33].

$$\text{Word rate} = \frac{\text{Number of words in the text}}{\text{Number of sentences in the text}}$$

$$\text{Difficulty ratio} = \frac{\text{Number of foreign words, idioms, terms concepts, methaphors, smiles, symbols, formulas in the text}}{\text{Number of words in the text}}$$

$$\text{Meaning ratio} = \frac{\text{Number of foreign words, idioms, terms concepts, methaphors, smiles, symbols, formulas in the text}}{\text{Number of sentences in the text}}$$

$$\text{Understandability rate} = \frac{\text{Meaning ratio}}{\text{Word rate}} \times \text{Difficulty ratio}$$

According to the Sönmez formula, the understandability rates and understandability levels are as follows: "0-0.00001: full communication is achieved", "0.00099-0.0001: text is clear and understandable", "0.03-0.001: text is understandable", "0.08-0.04: text can be understood with help", "0.15-0.09: text is difficult to understand", "0.25-0.16: text is blurry", "0.98-0.26: text is meaningless", "1.00-0.99: text is completely meaningless" [33].

Çetinkaya-Uzun Readability Formula

This formula was proposed by Çetinkaya [34] and is used to define and classify the readability levels of Turkish texts.

$$\text{Readability score (RS)} = 118.823 - (25.987 \times \text{AWL}) - (0.971 \times \text{ASL})$$

$$\text{Average Sentence Length (ASL)} = \frac{\text{Total number of words}}{\text{Total number of sentences}}$$

$$\text{Average Word Length (AWL)} = \frac{\text{Total number of syllables}}{\text{Total number of words}}$$

According to the Çetinkaya formula, readability levels are classified as "0-34: inadequate reading level", "35-50: educational reading level", "51+: independent reading level" [34].

Bezirci-Yılmaz Readability Formula

It is another formula proposed to define and classify the readability levels of Turkish texts [35]. Bezirci-Yılmaz formula has more detailed variables than the previous formulas. While the average word length is directly included in the equation in Ateşman and Çetinkaya-Uzun formulas, in Bezirci-Yılmaz formula, words are included in the equation separately according to the number of syllables.

$$\text{New Readability Value (NRV)} = \sqrt{(\text{AWS} \times [(H3 \times 0.84) + (H4 \times 1.5) + (H5 \times 3.5) + (H6 \times 26.35)])}$$

$$\text{Average number of words in a sentence (AWS)} = \frac{\text{Number of words in the text}}{\text{Number of sentences}}$$

$$H3 = \frac{\text{Number of three-syllable words in the text}}{\text{Total number of sentences}}$$

$$H4 = \frac{\text{Number of four-syllable words in the text}}{\text{Total number of sentences}}$$

$$H5 = \frac{\text{Number of five-syllable words in the text}}{\text{Total number of sentences}}$$

$$H6 = \frac{\text{Number of words with six or more syllables in the text}}{\text{Total number of sentences}}$$

According to the Bezirci-Yılmaz formula, readability levels are evaluated as "1-8: primary school", "9-12: high school", "13-16: undergraduate", "16+: academic" [35].

Statistical Analysis

Descriptive statistics of the data obtained from the study are given as mean±standard deviation or median (Q1-Q3) for quantitative variables and percentage values for categorical variables. One-Way Analysis of Variance (ANOVA) and Kruskal Wallis test with Dunn's posthoc test were used to compare the Ateşman Readability Formula, Sönmez Formula, Çetinkaya-Uzun Readability Formula and Bezirci-Yılmaz Readability Formula scores of the chatbots. The Fisher-Freeman-Halton test was used to compare the accuracy rates of the chatbots. In evaluating the accuracy rates of the answers recorded for each chatbot, the agreement between the experts was evaluated with the Kendall concordance coefficient. The Kendall concordance coefficient measures the amount of agreement between the decisions of K experts, each measured with an ordinal scale for N items [36]. The analyses were performed with IBM SPSS Statistics 23.0 program. $p<0.05$ was considered statistically significant.

Table 1. Questions Addressed to Chatbots

1.	In a city, the vitamin D levels in the blood of 35 women living in rural areas and 35 women living in urban areas were examined. It is known that the vitamin D levels in both groups follow a normal distribution. Which significance test should be used to evaluate whether there is a difference between these two regions in terms of the women's vitamin D levels?
2.	The hemoglobin levels of twenty anemic patients are measured before and one month after receiving an iron supplement. It is known that the hemoglobin levels do not follow a normal distribution in both measurements. Which significance test should be used to evaluate whether there is a difference in hemoglobin levels between the two measurements?
3.	The correlation coefficient between the ages and systolic blood pressures of a group of individuals is found to be $r = 0.80$ ($p<0.001$). Given that the variables follow a normal distribution, how is the relationship between these two variables interpreted?
4.	What does the normal distribution of data mean in significance tests? What are the most commonly used tests to evaluate the normal distribution of data?
5.	Can you interpret the given analysis outputs statistically? (IBM SPSS Statistics 23 output)
6.	According to the results of a previous similar study, the average of experimental group was found to be 25±2.3, while the average of control group was found to be 20±1.8. Can you calculate the minimum sample size required for the experimental and control groups in a study to be conducted on a similar topic ($\alpha=0.05$; $\beta=0.80$)?
7.	In a study, 15 out of 30 patients with headaches are given Drug A and 15 are given Drug B. The time taken for the drugs to relieve the pain is recorded. It is desired to assess whether there is a difference between Drug A and Drug B in terms of time taken to relieve pain. What is the hypothesis of such a study?
8.	When data do not follow a normal distribution, which measures of central tendency and dispersion should be used?
9.	Can you interpret the name and output of the given graph statistically? (IBM SPSS Statistics 23 output)
10.	Can you explain the concepts of "variable" and "parameter" in statistics?

RESULTS

The Ateşman readability formula average scores of the responses given to the questions specified in Table 2 of ChatGPT4, Gemini and Copilot chatbots were 42.88±7.77, 43.28±6.05 and 49.46±13.4, respectively. As a result of the Ateşman readability formula, it was concluded that the readability levels of all chatbots were "difficult". The understandability rates obtained from the Sönmez formula for the responses received from ChatGPT4, Gemini and Copilot chatbots were calculated as 0.003±0.003; 0.002±0.002 and 0.006±0.009, respectively, and "text is understandable" for all chatbots. Similarly, the Çetinkaya-Uzun readability formula averages of the responses given by ChatGPT4, Gemini and Copilot chatbots were calculated as 30.16±4.51, 29.39±3.74 and 33.77±7.71, respectively and "insufficient reading level" was found for all chatbots. Finally, the Bezirci-Yılmaz readability formula averages of the responses received from ChatGPT4, Gemini and Copilot chatbots were calculated as 47.0±22.92; 93.55±12.56 and 46.92±10.72, respectively, and it was concluded that the Bezirci-Yılmaz readability level for all chatbots was at the "academic level".

When artificial intelligence chatbots were compared in terms of readability and understandability measures, only Gemini chatbot ($p<0,001$ vs ChatGPT 4 and $p<0,001$ vs Copilot) showed a statistical difference from ChatGPT4 and Copilot chatbots in terms of Bezirci-Yılmaz readability level (Table 2).

No statistically significant difference was obtained in terms of correct answer rates of the artificial intelligence chatbots ($p=0.819$) (Table 3). The accuracy rates of ChatGPT4, Gemini and Copilot chatbots were found to be 60%, 60% and 80%, respectively. Also, the Kendall concordance coefficients of agreement between the experts for ChatGPT4, Gemini and Copilot chatbots were found 66%, 83,1% and 85,4%, respectively.

Table 2. Comparison of readability and understandability measures in chatbots

Readability and Comprehensibility Criteria	Chatbots	Mean±SD	Median (Q1-Q3)	p		
Ateşman Readability Formula	ChatGPT 4 (A)	42.88±7.77	42.93 (37.95-47.46)	0.266 Ψ		
	Gemini (B)	43.28±6.05	44.14 (37.97-45.86)			
	Copilot (C)	49.46±13.4	48.51 (42.64-55.16)			
Sönmez Formula	ChatGPT 4 (A)	0.0032±0.0032	0.0031 (0.0005-0.0049)	0.401¥		
	Gemini (B)	0.0023±0.0025	0.0019 (0.0002-0.0032)			
	Copilot (C)	0.0062±0.0086	0.0036 (0.0006-0.0071)			
Çetinkaya-Uzun Readability Formula	ChatGPT 4 (A)	30.16±4.51	31.33 (26.52-32.54)	0.194 Ψ		
	Gemini (B)	29.39±3.74	29.51 (26.28-32.65)			
	Copilot (C)	33.77±7.71	33.21 (26.48-37.25)			
Bezirci-Yılmaz Readability Formula	ChatGPT 4 (A)	47±22.92	42.78 (26.09-74.36)	**<0.001*¥ B>A=C**	ChatGPT 4 - Copilot	**0,859**
	Gemini (B)	93.55±12.56	96.37 (88.28-100.19)		ChatGPT 4 - Gemini	**<0,001***
	Copilot (C)	46.92±10.72	44.73 (39.91-53.46)		Copilot- Gemini	**<0,001***

*p<0.05; Ψ: One-way ANOVA; ¥: Kruskal Wallis test (Dunns' posthoc test); SD: Standard Deviation; Q1: 1st Quartile; Q3: 3rd Quartile

Table 3. Comparison of accuracy rates of chatbots in line with expert opinions

Variables n (%)		False n (%)	Partially True n (%)	True n (%)	p
Group	Chat GPT	1 (10)	3 (30)	6 (60)	0.819
	Gemini	2 (20)	2 (20)	6 (60)	
	Copilot	1 (10)	1 (10)	8 (80)	

The Fisher-Freeman-Halton test

DISCUSSION

According to the answers given to the questions addressed to the artificial intelligence chatbots, it was concluded that the score levels obtained from the Ateşman readability formula were "difficult". Similarly, it was determined that the score levels obtained from the Çetinkaya-Uzun readability formula were "insufficient reading level". The readability levels of the answers being "difficult" or "insufficient reading level" can be interpreted as the biostatistics literacy levels of the researchers using the chatbots should be high. Because the use of expressions and terms specific to the field of science reduces the readability levels of the answers. This shows that the researchers using the chatbots should have sufficient knowledge of field-specific expressions and terms. It was determined that the score levels obtained from the Bezirci-Yılmaz readability formula were "academic" level. Therefore, it can be said that in order for the answers given by the chatbots to be readable, the users should have an academic level of education specific to that field. It was also determined that the accuracy rates of the chatbots were not sufficient.

There are many studies evaluating the readability, understandability, and accuracy of artificial intelligence chatbots in different fields of health sciences with similar results to our study [28, 31, 32, 37-43]. Hancı et al. [28] examined ChatGPT, Bard, Gemini, Copilot, Perplexity chatbots using ARI (Automated Readability Index), FKG (Flesch-Kincaid Grade), and FRE (Flesch Reading Ease) indexes to evaluate the readability, reliability, and quality of responses related to palliative care. It was concluded that the quality and readability of the responses were not sufficient and that the responses provided by the chatbots were at the 6th grade reading level. Hershenhouse et al. [31] evaluated ChatGPT using Flesch-Kincaid Reading Ease (FRE), Flesch-Kincaid Grade Level (FKG), and Automated Readability Index (ARI) in their study to evaluate the level of prostate cancer knowledge and concluded that accuracy and understandability was low, and texts are readable. Önder et al. [32] evaluated ChatGPT 4.0 using FRE and FKG to evaluate the reliability and readability of responses related to hypothyroidism in pregnancy, and as a result of the FRE score, it was found that the text was difficult to read, and the level of education required to understand the responses was university level. Güven et al. [37] evaluated the performance of ChatGPT 3.5, ChatGPT 4.0 and Google Gemini artificial intelligence chatbots using FRE and FKG measures in responding to patient questions about dental injuries as a result of trauma and concluded that readability was difficult, and the level of education required university level reading skills. In the study conducted by Gajjar et al. [38], ChatGPT 3.0, ChatGPT 3.5 and ChatGPT 4.0 were evaluated using FRE and FKG measures to assess the accuracy of responses given to patients' questions for neurosurgical procedures. The study found that the readability level was difficult, and the education level was at a postgraduate level. In the study conducted by Ayo-Ajibola et al. [39], ChatGPT was evaluated using FRE and FKG measures on tracheostomy care recommendations. The readability of the answers to the questions corresponding to the low education level was found to be easy, the readability of the self-care questioning category was found to be difficult, and the readability levels for all questioning categories, except for the special situation questioning, were found to be at the 12th grade level or above. In the study conducted by Gondode et al. [40], they compared the accuracy of patient education tools for chronic pain medications created by ChatGPT with materials from traditional sources. Readability was evaluated using FRE and FKG measures. They concluded that traditional sources are more readable and potentially easier to understand. Steimetz et al. [41] examined Google Bard, ChatGPT using FRE and FKG to evaluate the ability to accurately explain pathology reports to patients and suggested that artificial intelligence chatbots can simplify pathology reports for patients and identify key details important for patient management, however, they concluded that interpretations should be used with caution as they are not perfect and that fact-checking solutions should be developed before integrating these tools into the healthcare environment. Carlson et al. [42] examined ChatGPT, Google Bard, Microsoft Bing, Perplexity, Claude using FRE and FKG to evaluate the accuracy and readability of responses to questions about vasectomy and concluded that all five artificial intelligence chatbots had an average FRE score below 50 and above a 10th grade reading level. They suggested that artificial intelligence chatbots may perform similarly in terms of their accuracy but may differ in terms of ease of understanding by the general public. Pradhan et al. [43] evaluated ChatGPT, DocsGPT, Google Bard, and Bing Chat using FRE and FKG to compare human-based patient education materials on cirrhosis, and concluded that the readability level was readable by someone with an 8th grade education level and understandable, but recommended that further work be done to easily accepted artificial intelligence chatbots in routine clinical practice.

Limitations and Strengths

In this study, parallel results were obtained with the results of similar studies. The limitation of the study can be stated as the small number of questions addressed to artificial intelligence chatbots. While indexes developed for foreign languages were used in previous studies evaluating Turkish responses in chatbots, the use of indexes adapted to Turkish in this study can be considered as a strength of the study. In addition, it is thought that being the first research in the field of biostatistics will contribute to the literature by providing an important perspective.

CONCLUSIONS

Recent studies have shown that the results of artificial intelligence chatbots are not readable, understandable, and have low accuracy levels. Although chatbots tend to provide accurate information, they have limitations available. One of the limitations of chatbots is the lack of evidence-based information sources. It can be difficult to know whether the information is reliable because it is not clear whether it is obtained from a valid source [28].

Although artificial intelligence chatbots are practical in terms of fast access to information and ease of use today, their readability, understandability and accuracy are not sufficient in areas that require expertise, such as biostatistics. Therefore, considering that biostatistics is an integral part of medical and health sciences research, researchers need to have a high level of knowledge, experience and academic reading level specific to the field of biostatistics. Although access to artificial intelligence chatbots is thought to be fast, easy and practical, it is seen that the information obtained from chatbots is not completely accurate and the information they provide is insufficient. As a result, it is considered appropriate that chatbots cannot be used to contribute to science, but only as a tool on the way to knowledge.

“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.” (R. A. Fischer, 1930).

Acknowledgments: This study was presented as an oral presentation at the 25th National and 8th International Biostatistics Congress held in Sakarya on 18-20 November 2024.

Conflict of interest: The authors declare that they have no conflicts of interest.

Funding: None.

Ethical Approval: No need.

Informed Consent: No need.

Author Contributions: İD: Conception, Design, Supervision, Materials, Analysis and/or Interpretation, Literature Review, Writing, Critical Review. PG: Design, Supervision, Materials, Analysis and/or Interpretation Literature Review, Writing, Critical Review. İB: Materials, Analysis and/or Interpretation, Literature Review, Writing. BİB: Materials, Data Collection and/or Processing, Literature Review, Writing

REFERENCES

[1] Turing AM (1950) Computing Machinery and Intelligence. Mind 59(236):433–460. https://doi.org/10.1093/mind/LIX.236.433

[2] McCarthy J, Minsky ML, Rochester N, Shannon CE (2006) A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. AI Mag. 27(4):12-14. https://doi.org/10.1609/aimag.v27i4.1904

[3] Samuel AL (1959) Some studies in machine learning using the game of checkers. IBM J Res Dev. 3(3):210-229. https://doi.org/10.1147/rd.33.0210

[4] Pirim AGH (2006) Artificial intelligence [Yapay Zeka]. Yaşar University E-Journal 1(1):81-93. ([In Turkish])

[5] Ozturk K, Sahin ME (2018) An overview of artificial neural networks and artificial intelligence [Yapay Sinir Ağları ve Yapay Zekâ’ya Genel Bir Bakış]. Takvim-i Vekayi 6(2):25-36. ([In Turkish])

[6] Lillicrap D, Morrissey JH (2023) Artificial intelligence, science, and learning. J Thromb Haemost. 21(4):709. https://doi.org/10.1016/j.jtha.2023.01.026

[7] Vedantham S, Shazeeb MS, Chiang A, Vijayaraghavan GR (2023) Artificial Intelligence in Breast X-Ray Imaging. Semin Ultrasound CT MR. 44(1):2–7. https://doi.org/10.1053/j.sult.2022.12.002

[8] Yoon C, Jones K, Goker B, Sterman J, Mardakhaev E (2025) Artificial Intelligence Applications in MR Imaging of the Hip. Magn Reson Imaging Clin N Am. 33(1):9–18. https://doi.org/10.1016/j.mric.2024.05.003

[9] Huang S, Yang J, Shen N, Xu Q, Zhao Q (2023) Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. Semin Cancer Biol. 89:30–37. https://doi.org/10.1016/j.semcancer.2023.01.006

[10] Lotter W, Hassett MJ, Schultz N, Kehl KL, Van Allen EM, Cerami E (2024) Artificial Intelligence in Oncology: Current Landscape, Challenges, and Future Directions. Cancer Discov. 14(5):711–726. https://doi.org/10.1158/2159-8290.CD-23-1199

[11] Itchhaporia D (2022) Artificial intelligence in cardiology. Trends Cardiovasc Med. 32(1):34–41. https://doi.org/10.1016/j.tcm.2020.11.007

[12] Miller RJH (2023) Artificial Intelligence in Nuclear Cardiology. Cardiol Clin. 41(2):151–161. https://doi.org/10.1016/j.ccl.2023.01.004

[13] Jacobson BC (2023) The Use of Artificial Intelligence in Gastroenterology: A Glimpse Into the Present. Clin Transl Gastroenterol. 14(10):e00653. https://doi.org/10.14309/ctg.0000000000000653

[14] Ahmed T, Rabinowitz LG, Rodman A, Berzin TM (2024) Generative Artificial Intelligence Tools in Gastroenterology Training. Clin Gastroenterol Hepatol. 22(10):1975–1978. https://doi.org/10.1016/j.cgh.2024.05.050

[15] Srivastava O, Tennant M, Grewal P, Rubin U, Seamone M (2023) Artificial intelligence and machine learning in ophthalmology: A review. Indian J Ophthalmol. 71(1):11–17. https://doi.org/10.4103/ijo.IJO_1569_22

[16] Honavar SG (2022) Artificial intelligence in ophthalmology - Machines think!. Indian J Ophthalmol. 70(4):1075–1079. https://doi.org/10.4103/ijo.ijo_644_22

[17] Scheer JK, Ames CP (2024) Artificial Intelligence in Spine Surgery. Neurosurg Clin N Am. 35(2):253–262. https://doi.org/10.1016/j.nec.2023.11.001

[18] Benzakour A, Altsitzioglou P, Lemée JM, Ahmad A, Mavrogenis AF, Benzakour T (2023) Artificial intelligence in spine surgery. Int Orthop. 47(2):457–465. https://doi.org/10.1007/s00264-022-05517-8

[19] Eric A, Ozgur EG, Asker OF, Bekiroglu N (2024) ChatGPT and its Use in Health Sciences. CBU-SBED 11(1):176-182. https://doi.org/10.34087/cbusbed.1262811

[20] Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke F (2024) Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: A pilot study. J Dent. 144:104938. https://doi.org/10.1016/j.jdent.2024.104938

[21] Issaiy M, Zarei D, Saghazadeh A (2023) Artificial Intelligence and Acute Appendicitis: A Systematic Review of Diagnostic and Prognostic Models. World J Emerg Surg. 18(1):59. https://doi.org/10.1186/s13017-023-00527-2

[22] Gore JC (2020) Artificial intelligence in medical imaging. Magn Reson Imaging. 68:A1–A4. https://doi.org/10.1016/j.mri.2019.12.006

[23] Kim ES, Eun SJ, Kim KH (2023) Artificial Intelligence-Based Patient Monitoring System for Medical Support. Int Neurourol J. 27(4):280–286. https://doi.org/10.5213/inj.2346338.169

[24] Smith A, Arena R, Bacon SL, Faghy MA, Grazzi G, Raisi A, Vermeesch AL, Ong'wen M, Popovic D, Pronk NP (2024) Recommendations on the use of artificial intelligence in health promotion. Prog Cardiovasc Dis. 87:37-43. https://doi.org/10.1016/j.pcad.2024.10.003

[25] Zhao T, Cui J, Hu J, Dai Y, Zhou Y (2022) Is Artificial Intelligence Customer Service Satisfactory? Insights Based on Microblog Data and User Interviews. Cyberpsychol Behav Soc Netw. 25(2):110–117. https://doi.org/10.1089/cyber.2021.0155

[26] Bawack RE, Wamba SF, Carillo KDA, Akter S (2022) Artificial intelligence in E-Commerce: a bibliometric study and literature review. Electron Mark. 32(1):297–338. https://doi.org/10.1007/s12525-022-00537-z

[27] Mohammadi SS, Khatri A, Jain T, Thng ZX, Yoo WS, Yavari N, Bazojoo V, Mobasserian A, Akhavanrezayat A, Tuong Than NT, Elaraby O, Ganbold B, El Feky D, Nguyen BT, Yasar C, Gupta A, Hung JH, Nguyen QD (2024) Evaluation of the Appropriateness and Readability of ChatGPT-4 Responses to Patient Queries on Uveitis. Ophthalmol Sci. 5(1):100594. https://doi.org/10.1016/j.

xops.2024.100594

[28] Hancı V, Ergün B, Gül Ş, Uzun Ö, Erdemir İ, Hancı FB (2024) Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. Medicine 103(33):e39305. https://doi.org/10.1097/MD.0000000000039305

[29] Golan R, Ripps SJ, Reddy R, Loloi J, Bernstein AP, Connelly ZM, Golan NS, Ramasamy R (2023) ChatGPT's Ability to Assess Quality and Readability of Online Medical Information: Evidence From a Cross-Sectional Study. Cureus 15(7):e42214. https://doi.org/10.7759/cureus.42214

[30] Gibson D, Jackson S, Shanmugasundaram R, Seth I, Siu A, Ahmadi N, Kam J, Mehan N, Thanigasalam R, Jeffery N, Patel MI, Leslie S (2024) Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: Multimetric Assessment. J Med Internet Res. 26:e55939. https://doi.org/10.2196/55939

[31] Hershenhouse JS, Mokhtar D, Eppler MB, Rodler S, Storino Ramacciotti L, Ganjavi C, Hom B, Davis R J, Tran J, Russo GI, Cocci A, Abreu A, Gill I, Desai M, Cacciamani GE (2024) Accuracy, readability, and understandability of large language models for prostate cancer information to the public. Prostate Cancer Prostatic Dis. https://doi.org/10.1038/s41391-024-00826-y

[32] Onder C, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz S (2024) Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep. 14:243. https://doi.org/10.1038/s41598-023-50884-w

[33] Kalyoncu MR, Memiş M (2024) Comparison of Readability Formulas Created and Consistency Query for Turkish [Türkçe İçin Oluşturulmuş Okunabilirlik Formüllerinin Karşılaştırılması ve Tutarlılık Sorgusu]. Journal of Mother Tongue Education 12:417-436. ([In Turkish]) https://doi.org/10.16916/aded.1434650

[34] Çetinkaya G (2010) Identification and classification of readability levels of Turkish texts (Unpublished Doctoral Thesis)[Türkçe Metinlerin Okunabilirlik Düzeylerinin Tanimlanmasi ve Siniflandirilmasi]. Ankara University, Ankara. ([In Turkish]).

[35] Bezirci B, Yılmaz AE (2010) A software library for measuring the readability of texts and a new readability criterion for Turkish [Metinlerin Okunabilirliğinin Ölçülmesi Üzerine Bir Yazilim Kütüphanesi Ve Türkçe İçin Yeni Bir Okunabilirlik Ölçütü]. DEUFMD. 12(3):49-62. ([In Turkish]).

[36] Doğan İ, Doğan N (2014) Adım adım çözümlü parametrik olmayan istatistiksel yöntemler, 1st edn. Detay Yayıncılık, Ankara.

[37] Guven Y, Ozdemir OT, Kavan MY (2024) Performance of Artificial Intelligence Chatbots in Responding to Patient Queries Related to Traumatic Dental Injuries: A Comparative Study. Dent Traumatol. https://doi.org/10.1111/edt.13020

[38] Gajjar AA, Kumar RP, Paliwoda ED, Kuo CC, Adida S, Legarreta AD, Deng H, Anand SK, Hamilton DK, Buell TJ, Agarwal N, Gerszten PC, Hudson JS (2024) Usefulness and Accuracy of Artificial Intelligence Chatbot Responses to Patient Questions for Neurosurgical Procedures. Neurosurgery. https://doi.org/10.1227/neu.0000000000002856

[39] Ayo-Ajibola O, Davis RJ, Lin ME, Vukkadala N, O'Dell K, Swanson MS, Johns MM 3rd, Shuman EA (2024) TrachGPT: Appraisal of tracheostomy care recommendations from an artificial intelligent Chatbot. Laryngoscope Investig Otolaryngol. 9(4):e1300. https://doi.org/10.1002/lio2.1300

[40] Gondode P, Duggal S, Garg N, Sethupathy S, Asai O, Lohakare P (2024) Comparing patient education tools for chronic pain medications: Artificial intelligence chatbot versus traditional patient information leaflets. Indian J Anaesth. 68(7):631–636. https://doi.org/10.4103/ija.ija_204_24

[41] Steimetz E, Minkowitz J, Gabutan EC, Ngichabe J, Attia H, Hershkop M, Ozay F, Hanna M G, Gupta R (2024) Use of Artificial Intelligence Chatbots in Interpretation of Pathology Reports. JAMA Netw Open. 7(5):e2412767. https://doi.org/10.1001/jamanetworkopen.2024.12767

[42] Carlson JA, Cheng RZ, Lange A, Nagalakshmi N, Rabets J, Shah T, Sindhwani P (2024) Accuracy and Readability of Artificial Intelligence Chatbot Responses to Vasectomy-Related Questions: Public Beware. Cureus 16(8):e67996. https://doi.org/10.7759/cureus.67996

[43] Pradhan F, Fiedler A, Samson K, Olivera-Martinez M, Manatsathit W, Peeraphatdit T (2024) Artificial intelligence compared with human-derived patient educational materials on cirrhosis. Hepatol Commun. 8(3):e0367. https://doi.org/10.1097/HC9.0000000000000367

How to Cite;

Dogan I, Gunel P, Berk I, Berk IB (2024) Evaluation of the Readability, Understandability, and Accuracy of Artificial Intelligence Chatbots in Terms of Biostatistics Literacy. Eur J Ther. 30(6):900-909. https://doi.org/10.58600/eurjther2569