Tıp Öğrencilerinin Biyoistatistik Sınavında ChatGPT-3.5 ve ChatGPT-4 Performanslarının Karşılaştırılması: Bir Eğitim Asistanı Olarak Artıları ve Eksileri: Kesitsel Çalışma

***

Comparing the Performance of Medical Students, ChatGPT-3.5 and ChatGPT-4 in Biostatistics Exam: Pros and Cons as an Education Assistant: A Cross-Sectional Research

Ömer Faruk ASKER[1]
Emrah Gökay ÖZGÜR[2]
Alper ERİÇ[3]
Nural BEKİROĞLU[4]

DOI:10.33461/uybisbbd.1329650

Öz

Makale Bilgileri

Makale Türü:
Araştırma Makalesi

Geliş Tarihi:
19.07.2023

Kabul Tarihi:
16.09.2023

Araştırmalar, tıp öğrencilerinin biyoistatistik konusundaki bilgi düzeylerinin beklenenden düşük olduğunu göstermiştir. Bu durum biyoistatistik eğitiminde yeni yöntemlerin uygulanması ihtiyacını doğurmaktadır. Bu çalışmanın amacı, ChatGPT'nin biyoistatistik alanında bir eğitim asistanı olarak uygulanabilirliğini değerlendirmektir. ChatGPT, OpenAI tarafından geliştirilmiş bir doğal dil işleme modelidir. Kullanıcılar tarafından sorulan sorulara insan benzeri cevaplar vermekte ve bilgi edinmek için çeşitli alanlarda kullanılmaktadır. ChatGPT, en yeni GPT-4 modeliyle çalışırken, önceki sürüm olan GPT-3.5 halen kullanımdadır. Bu çalışmada da 245 Marmara Üniversitesi Tıp Fakültesi öğrencisinin biyoistatistik performansları, temel biyoistatistik konularını kapsayan bir sınav kullanılarak ChatGPT-3.5 ve ChatGPT-4 ile karşılaştırıldı. SonuçlarElde edilen bulgulara göre ChatGPT-3.5 sınavda %80, ChatGPT-4 ise %100 başarı oranı elde etmiştir. Buna karşılık, öğrenciler %67,9 başarı oranı elde ettiler. Ayrıca ChatGPT-3.5 matematiksel hesaplama gerektiren sorularda sadece %33 başarı oranı kaydederken, ChatGPT-4 bu sorularda %100 başarı oranı elde etmiştir. Sonuç olarak ChatGPT, biyoistatistik alanında potansiyel bir eğitim asistanıdır. Mevcut sürümdeki başarısı önceki sürüme göre önemli ölçüde artmıştır. Yeni sürümler çıktıkça daha fazla çalışmaya ihtiyaç duyulacaktır.

***Anahtar Kelimeler:** ChatGPT, Biyoistatistik, Eğitim, NLP.*

Abstract

Article Info

Paper Type:
Research Paper

Received:
19.07.2023

Accepted:
16.09.2023

Studies have shown that the level of knowledge in biostatistics among medical students is lower than expected. This situation calls for the need to implement new methods in biostatistics education. The aim of this study is to evaluate the feasibility of ChatGPT as an education assistant in biostatistics. ChatGPT is a natural language processing model developed by OpenAI. It provides human-like responses to questions asked by users and is utilized in various fields for gaining information. ChatGPT operates with the latest GPT-4 model, while the previous version, GPT-3.5, is still in use. In this study the biostatistics performance of 245 Marmara University School of Medicine students was compared to ChatGPT-3.5 and ChatGPT-4 using an exam covering basic biostatistics topics. According to findings, ChatGPT-3.5 achieved 80% success rate in the exam, while ChatGPT-4 achieved 100% success rate. In contrast, the students achieved 67.9% success rate. Furthermore, ChatGPT-3.5 only recorded 33% success rate in questions requiring mathematical calculations, while ChatGPT-4 achieved 100% success rate in these questions. In conclusion, ChatGPT is a potential education assistant in biostatistics. Its success has increased significantly in the current version compared to the previous one. Further studies will be needed as new versions are released.

***Keywords:** ChatGPT, Biostatistics, Education, NLP.*

Atıf/ to Cite (APA): Asker, Ö. F., Özgür, E. G., Eriç, A. & Bekiroğlu, N. (2023). Comparing the Performance of Medical Students, ChatGPT-3.5 and ChatGPT-4 in Biostatistics Exam: Pros and Cons as an Education Assistant: A Cross-Sectional Research. International Journal of Management Information Systems and Computer Science, 7(2), 85-94. DOI: 10.33461/uybisbbd.1329650

[1] omrfrkaskr@hotmail.com

[2] Asst. Prof., Faculty of Medicine, Marmara University, emrahgokayozgur@gmail.com

[3] alper.eric10@gmail.com

[4] Prof. Dr. Faculty of Medicine, Marmara University, nural@marmara.edu.tr

1. INTRODUCTION

Chat Generative Pre-Trained Transformer (ChatGPT) is an artificial intelligence language model developed by OpenAI that was released on November 30, 2022, as a product of the natural language processing (NLP) subfield of artificial intelligence. Unlike other artificial intelligence models, ChatGPT has been trained with various databases to answer user questions, but it can also respond to consecutive questions, accept, and correct errors in its responses, and refuse to answer inappropriate questions. These properties provide users with a human-like conversation experience. Therefore, ChatGPT's usage has become widespread reaching one million users just five days after its release. 1

It has been working on chatbots for approximately 80 years. Based on the possibility of machines being able to think, humanity has made many chatbot attempts until today. All these chatbots work relying on a system called natural language processing (NLP), which aims to enable machines to understand human language. 2 The latest version of ChatGPT is ChatGPT-4, previous model ChatGPT-3.5 is also available. On March 14, 2023, OpenAI released ChatGPT-4, almost 1 year after ChatGPT-3.5. According to OpenAI, in comparison to GPT-3.5, GPT-4 has an 82% reduced likelihood of replying to queries involving prohibited content. Additionally, GPT-4 exhibits a 40% higher probability of generating accurate answers.3,4

There are numerous studies related to the capabilities of ChatGPT-3.5. Examples include its usability in medical education, 5,6,7,8, interpreting radiology reports 9 and its abilities in mathematics. 10 In contrast, capability studies on ChatGPT-4 are still limited in number. According to information released by OpenAI, ChatGPT-4, which was tested in various fields such as statistics, mathematics, history, and biology, has achieved much higher scores compared to ChatGPT-3.5. 3,4

Biostatistics is a scientific discipline that deals with the application and development of statistical theory and methods in the field of life and health sciences.11 A medical professional must have sufficient knowledge of biostatistics to understand research in the medical literature, interpret statistical results, and increase his/her utilization of the literature. This competency is examined through the concept of Biostatistics.12 The evidence-based medicine, which began to be used in the late 20th century, highlights the need for medical knowledge to be produced based on scientific study and statistically proven data, rather than solely relying on the individual experience and preferences of expert clinicians. In this regard, medical professionals need to have a strong knowledge of biostatistics to use evidence-based medical information.13

There are few studies in the literature that evaluate the level of biostatistical knowledge and literacy of pre-graduate and post-graduate medical students. A literature search was conducted on the Google Scholar database using the keyword groups "biostatistical knowledge" and "medical students" for studies published from 2019 to the present day, and six studies evaluating the biostatistical knowledge levels of students were reviewed. Two of these studies have evaluated the knowledge levels of residents, 14,15 while four have evaluated the knowledge levels of medical students currently enrolled in undergraduate programs. 16,17,18,19 All studies reported that the biostatistical knowledge and literacy levels of students were lower than expected. In this context, some academics have expressed concerns that biostatistical knowledge is not being used accurately enough in published articles, including high-impact factor journals, and that this lack of knowledge leads to the use of incorrect statistical methods.20,21 Therefore, the importance of new learning techniques and programs in biostatistics education has been emphasized.

The success of ChatGPT in various fields, including medicine, has been examined in different studies by applying the same examinations used for students to ChatGPT. For example, in a study conducted in Turkey, the performance of ChatGPT in the field of Anatomy was evaluated using an examination administered to students, and ChatGPT was found to outperform the students.22 However, there is no study available that compares ChatGPT's 3.5 and 4.0 versions or evaluates its performance in the field of biostatistics.

The aim of this study is to compare the success rates of biostatistics questions in the committee exam for Marmara University Medical School students with both ChatGPT-3.5 and ChatGPT-4.5. Thus, the following evaluations will be provided regarding ChatGPT:

- Usability of its responses about biostatistics,
- Its ability to act as an assistant in biostatistics education,
- The competency difference between ChatGPT-3.5 and ChatGPT-4 in the field of biostatistics.

2. MATERIAL AND METHOD

In this study, the questions in the first year second term committee exam of Marmara University Medical School were used. 10 biostatistics questions were asked totally in the committee exam. The questions were prepared in English by the faculty members of the Biostatistics Department of Marmara University Medical School. The topics of biostatistics questions were asked in the committee exam were given below as; Principles of Statistical Analysis, Elements of Statistical Inference, Bayes' Theorem, Sampling, Distribution and Estimation, T-Test, Testing Statistical Hypothesis, Types of Errors in Statistical Inference, Probability and Probability Distribution, Parametric and Nonparametric Methods and Introduction to Statistical Analysis.

A total of 245 students participated in the exam, which was performed face-to-face at Marmara University Faculty of Medical School on January 19, 2023.The students' rates of answering the questions correctly were obtained from a software called Corporate Education Management and Planning System (Kurumsal Eğitim Yönetimi ve Planlama Sistemi - KEYPS). KEYPS is a software that provides assessment and evaluation services to various higher education institutions in Turkey, including Marmara University Medical School.23 After the committee exam, analyses related to the exam are published on the website of KEYPS. The rates of students' answering correctly to each biostatistics question asked in the exam, were obtained.

Each biostatistics question in the exam has been presented to ChatGPT-3.5 and ChatGPT-4 without any modification on February 28, 2023, and March 18, 2023, respectively. The responses and accuracy status provided by ChatGPT were recorded.

3. RESULTS

Table 1 displays the topics of the exam with the performance of the students, the performance of ChatGPT-3.5, the performance of ChatGPT-4. As examples of ChatGPT's responses, the first 6 questions of the exam and the answers provided by ChatGPT-3.5 and ChatGPT-4 are shown in Table 2 and Table 3.

In the exam consisting of a total of 10 biostatistics questions, ChatGPT-3.5 answered correctly 8 of these questions, achieving 80% success but ChatGPT-4 answered correctly all the questions, achieving 100% success. However, the average success rate per question for the students was found to be 67.9% (Table 1).

Table 1. Topics of questions and performances of students*, ChatGPT-3.5 and ChatGPT-4 in exam

Question	Topic	Students	ChatGPT-3.5	ChatGPT-4
1	Probability and Probability Distribution	60.73%	True	True
2	Bayes’ Theorem	79.35%	False	True
3	Introduction to Statistical Analysis	46.15%	True	True

4	Sampling, Distribution and Estimation	72.47%	True	True
5	Elements of Statistical Interference	87.04%	False	True
6	Testing Statistical Hypothesis	67.61%	True	True
7	Types of Errors in Statistical Inference	62.75%	True	True
8	Parametric and Nonparametric Methods	47.37%	True	True
9	T-Test	68.42%	True	True
10	Principles of Statistical Analysis	87.04%	True	True
	AVERAGE SCORE	**67.9%**	**80%**	**100%**

*Given values for students are average scores of all 245 students.

There was one question per topic, the success rates of the students according to the topics were found to be as follows: Principles of Statistical Analysis 87.04%, Elements of Statistical Inference 87.04%, Bayes' Theorem 79.35%, Sampling, Distribution and Estimation 72.47%, T-Test 68.42%, Testing Statistical Hypothesis 67.61%, Types of Errors in Statistical Inference 62.75%, Probability and Probability Distribution 60.73%, Parametric and Nonparametric Methods 47.37%, and Introduction to Statistical Analysis 46.15% (Table 1).

Out of the 10 questions in the exam, 3 of them required mathematical calculations (Table 2 and Table 3). ChatGPT-3.5 correctly answered only 1 of these 3 questions, resulting in a success rate of 33%. In contrast, ChatGPT-4 correctly answered all questions requiring mathematical calculations, achieving a 100% success rate. The average success rate of students in questions requiring mathematical calculations was found to be 70.8%.

The topics for which ChatGPT-3.5 provided correct answers in the exam are as follows: Probability and Probability Distribution, Introduction to Statistical Analysis, Sampling, Distribution and Estimation, Types of Errors in Statistical Inference, Parametric and Nonparametric Methods, T-Test and Principles of Statistical Analysis. However, topics which ChatGPT-3.5 provided wrong answers are Bayes’ Theorem and Elements of Statistical Inference. It was observed that the students gave correct answers at a high rate in the questions where ChatGPT-3.5 answered incorrectly. In contrast, ChatGPT-4 has accurately answered all the questions posed to it and has shown success in all topics.

Table 2. Questions 1-3, responses of ChatGPT-3.5 and ChatGPT-4 for these questions

Question*	Responses of ChatGPT-3.5	Responses of ChatGPT-4
1) Which of the information given about the binomial distribution is/are correct? I. The experiment consists of n identical trials II. The trials are independent III. The probability of success on a single trial is equal to p but varies from trial to trial A) Only I B) Only II C) Only III **D) I and II** E) II and III	The correct answer is D) I and II. The binomial distribution is a discrete probability distribution that describes the number of successes in a fixed number of independent trials, each with the same probability of success. The information given in options I and II are correct. Option I state that the experiment consists of n identical trials. This is true since the binomial distribution applies to situations where there are a fixed number of trials, each of which has the same probability of success. Option II states that the trials are independent. This is also true since the outcome of each trial does not affect the outcome of any other trial. Option III, on the other hand, is not correct. The probability of success is fixed and constant for all the trials in a binomial distribution. It does not vary from trial to trial.	D) I and II, The binomial distribution has the following characteristics: I. The experiment consists of n identical trials. II. The trials are independent. III. The probability of success on a single trial is equal to p and remains constant from trial to trial. Option III is incorrect because the probability of success does not vary from trial to trial in a binomial distribution.
2) In a neighborhood, 80% children were falling sick due flu and 20% due to measles and no other disease. The probability of observing rashes for measles is 0.80 and for flu is 0.05. If a child develops rashes, find the child's probability of having flu. A) 0.10 B) 0.15 **C) 0.20** D) 0.30 E) 0.40	Let's use Bayes' theorem to solve this problem. Let F be the event that the child has flu, and M be the event that the child has measles. We want to find the probability of having flu given that the child has rashes, i.e., P(F | R). By Bayes' theorem, we have: P(F | R) = P(R | F) P(F) / [P(R | F) P(F) + P(R | M) P(M)] where P(R | F) = 0.05 is the probability of observing rashes given the child has flu, and P(R | M) = 0.80 is the probability of observing rashes given the child has measles. Also, we know that P(F) = 0.80 is the prior probability of having flu, and P(M) = 0.20 is the prior probability of having measles. Substituting the values, we get: P(F | R) = 0.05 x 0.80 / [0.05 x 0.80 + 0.80 x 0.20] = 0.0667 Therefore, the probability of having flu given that the child has rashes is 0.0667, which is approximately 0.07. Therefore, the answer is not one of the given options.	We can use Bayes' theorem to solve this problem. Let F represent flu, M represent measles, and R represent rashes. We want to find P(F|R), which is the probability of having flu given that the child has rashes. Bayes' theorem states: P(F|R) = P(R|F) * P(F) / (P(R|F) * P(F) + P(R|M) * P(M)) We are given the following probabilities: P(F) = 0.80 (probability of flu) P(M) = 0.20 (probability of measles) P(R|F) = 0.05 (probability of rashes given flu) P(R|M) = 0.80 (probability of rashes given measles) Now, substitute the values into the formula: P(F|R) = (0.05 * 0.80) / (0.05 * 0.80 + 0.80 * 0.20) P(F|R) = 0.04 / (0.04 + 0.16) P(F|R) = 0.04 / 0.20 P(F|R) = 0.20 The correct answer is: C) 0.20
3) The average systolic blood pressure of a group of people is 130 with a standard deviation of 16. What is the standardized score (or z- score) of someone with an systolic blood pressure of 122? A) 1 B) -2 C) 4 **D) -0.5** E) -6	To calculate the z-score, we use the formula: z = (x - μ) / σ, where x is the individual's systolic blood pressure, μ is the population mean systolic blood pressure, and σ is the population standard deviation. Substituting the given values: z = (122 - 130) / 16, z = -0.5 Therefore, the standardized score (or z-score) of someone with a systolic blood pressure of 122 is -0.5. So, the answer is option D) -0.5.	To calculate the z-score, we use the formula: z = (X - μ) / σ where X is the individual's systolic blood pressure, μ is the average systolic blood pressure, and σ is the standard deviation of systolic blood pressure. In this case, X = 122, μ = 130, and σ = 16. z = (122 - 130) / 16 z = (-8) / 16 z = -0.5 The z-score for someone with a systolic blood pressure of 122 is: D) -0.5

*True answers of questions are **bold**.

Table 3. Questions 4-6, responses of ChatGPT-3.5 and ChatGPT-4 for these questions

Question*	Responses of ChatGPT-3.5	Responses of ChatGPT-4
4) The theoretical probability distribution of any statistic based on samples of size n drawn at random from a population is referred to as ………………… A) Binomial Distribution B) Poisson Distribution **C) Sampling Distribution** D) Bernoulli Distribution E) F-Distribution	The theoretical probability distribution of any statistic based on samples of size n drawn at random from a population is referred to as the "Sampling Distribution". Therefore, the correct answer is C) Sampling Distribution. The binomial distribution applies to situations where there are a fixed number of trials, each of which has the same probability of success. The Poisson distribution is used to model the number of occurrences of rare events. The Bernoulli distribution is a special case of the binomial distribution, which applies when there is only one trial. The F-distribution is used to test the equality of variances of two populations based on sample data.	The theoretical probability distribution of any statistic based on samples of size n drawn at random from a population is referred to as: C) Sampling Distribution
5) Which of the following p values is obtained, the H0 hypothesis is accepted? A) 0.03 B) 0.028 C) 0.001 **D) 0.059** E) 0.044	The decision to accept or reject the null hypothesis (H0) based on the p-value depends on the pre-determined significance level (α), which is usually set to 0.05 (5%) in most fields. If the p-value is less than or equal to the significance level ($p \leq \alpha$), then we reject the null hypothesis and conclude that the alternative hypothesis (H1) is supported. Conversely, if the p-value is greater than the significance level ($p > \alpha$), then we fail to reject the null hypothesis and conclude that there is not enough evidence to support the alternative hypothesis. Therefore, based on a significance level of 0.05, we would accept the null hypothesis for options D) 0.059 and E) 0.044, since their p-values are greater than 0.05. We would reject the null hypothesis for options A) 0.03, B) 0.028, and C) 0.001, since their p-values are less than or equal to 0.05. In summary: Accept H0 for D) 0.059 and E) 0.04, Reject H0 for A) 0.03, B) 0.028, and C) 0.001	It is important to note that we do not "accept" the null hypothesis (H0); instead, we "fail to reject" it. The p value is compared to a significance level (commonly denoted as α), which is usually set at 0.05. If the p-value is greater than the significance level, we fail to reject the null hypothesis. In this case, only one p-value is greater than 0.05: D) 0.059
6) A researcher wants to evaluate whether there is a difference in the amount of cigarette smoking depending on whether a person is diagnosed with major depression. A total of 100 people with and without depression were included in the study. He determined the amount of smoking of these people and analyzed them. As a result of the statistical analysis, the p value was found to be 0.044. **A) Reject H0** B) Fail to accept H1 C) Accept H0 D) Accept H1 E) Fail to reject H1	The p-value is a measure of the evidence against the null hypothesis (H0). A p-value of 0.044 means that there is a 4.4% chance of obtaining the observed results or more extreme results, assuming that the null hypothesis is true. The null hypothesis in this case is that there is no difference in the amount of cigarette smoking between people with and without major depression. The alternative hypothesis (H1) is that there is a difference in smoking behavior between these two groups. The conventional significance level is 0.05. If the p-value is less than or equal to the significance level, then we reject the null hypothesis; otherwise, we fail to reject the null hypothesis. Since the p-value of 0.044 is less than the significance level of 0.05, we can reject the null hypothesis and conclude that there is a statistically significant difference in the amount of cigarette smoking between people with and without major depression. Therefore, the correct answer is A) Reject H0.	The p-value (0.044) is less than the significance level of 0.05, which indicates that there is enough evidence to reject the null hypothesis (H0) in favor of the alternative hypothesis (H1). The correct answer is: A) Reject H0

*True answers of questions are **bold**.

4. CONCLUSION

Due to its training with a significant amount of data and providing fast and tailor-made responses to user questions, ChatGPT quickly reached many users. Its ability to present relevant information in a dialog format has led to investigations into its effectiveness in various fields, especially for ChatGPT-3.5. The usability and success of ChatGPT in medical education and national medical exams have been observed by researchers.1,5,6,7,8 A study conducted in Pakistan stated that ChatGPT could be used effectively in medical education, medical research, and clinical management due to its ability to provide learning assistance, personalized education with automatic grading.6 The success of ChatGPT in the USMLE exam has been examined in three different studies in the literature. According to the results of the studies, the exam scores obtained by ChatGPT and the quality of the answers it generates in response to questions indicate that it can be used as an efficient assistant in medical education.6,8,24

According to our study results, ChatGPT-3.5 answered correctly 8 out of the 10 questions. These two questions that ChatGPT-3.5 answered incorrectly, were the second and the fifth questions, which had success rates of 79.35% and 87.04% among students, respectively. Interestingly, these two questions were among the ones with the highest success rates among students. On the other hand, ChatGPT-4 gave correct answers for all the biostatistics questions. Therefore, it seems that the performance of ChatGPT-4 looks better than the previous versions.

ChatGPT-3.5 does not only provide the correct answers to the questions but also gives very helpful explanations even for the questions it answered incorrectly. For example, in first question of the exam (Table 2), ChatGPT-3.5 not only evaluated the veracity of the provided information but also elucidated the reasons for the correctness of the first and second pieces of information and the inaccuracy of the third piece of information. Similarly, in third question (Table 2), which has the lowest average success rate among students, ChatGPT-3.5 has provided the necessary formula and explicated the values in the formula according to the scenario presented in the question. ChatGPT-4 was found to be fully successful, and its explanations were shorter and more informative than ChatGPT-3.5's explanations.

Out of the 10 Biostatistics questions in the Committee exam we used in our study, three were numerical questions that required calculations, while the remaining seven were questions that required interpretation based on knowledge. ChatGPT-3.5 did not make any errors in the interpretation-based questions, but it answered two out of the three numerical questions incorrectly. In a study conducted at the University of Minnesota Law School, 25 ChatGPT-3.5 was presented with both mathematical reasoning and non-mathematically reasoning questions, and it was reported that ChatGPT-3.5 correctly answered 16 out of 31 (51.6%) non-mathematically reasoning questions. In our study, ChatGPT-3.5's success rate in these types of questions was found to be 100%. Additionally, in the Minnesota study, it was reported that ChatGPT-3.5 correctly answered 8 out of 29 (27.5%) mathematically reasoning questions, while in our study, the success rate for similar types of questions was found to be 33.3%. In both studies, ChatGPT's performance on questions requiring mathematical computation was found to be lower. Furthermore, in a study conducted on ChatGPT's performance in mathematics, it was reported that ChatGPT scored lower than an average mathematics graduate.10 Similarly, in our study, while the average success rate of students in questions requiring mathematical operations was 70.8%, ChatGPT-3.5's success rate was found to be 33.3%. However, ChatGPT-4's success rate was found to be 100%.

According to the results of our study, ChatGPT-4 has demonstrated full success in the basic biostatistics exam. It has a higher performance compared to ChatGPT-3.5 in terms of both the accuracy of the answers given to the questions and the explanatory nature of the answers. In research published by OpenAI, ChatGPT-3.5 and ChatGPT-4 were evaluated in the Advanced Placement (AP) Statistics exam. Similar to our findings, according to OpenAI's research, while ChatGPT-3.5 achieved a 40% success rate, ChatGPT-4 achieved more than 80% success.3,4

ChatGPT-3.5 answered the 2nd and 5th questions incorrectly (Table 2). In the 2nd question related to Bayes' theorem, ChatGPT-3.5 recognized that Bayes' theorem should be used even though it was not explicitly mentioned in the question, provided the appropriate formulas, and used the correct numbers in the formula. However, ChatGPT-3.5 made an error in the calculation and gave a result of 0.0667 instead of the correct answer of 0.2 for the operation "P (F | R) = 0.05 x 0.80 / [0.05 x 0.80 + 0.80 x 0.20]'. It seems that ChatGPT-3.5 understands the question correctly but makes a mistake in the calculation. Similarly, in an investigated study of ChatGPT-3.5's success in mathematics, it is reported that ChatGPT understands the question very well but solves it incorrectly.10

Another question that ChatGPT-3.5 answered incorrectly was the fifth question (Table 2). In this question, the significance level or the confidence interval was not specified because the students should know that the significance level is 0.05 at maximum, and this is emphasized many times to the students who took the exam in the relevant courses. Therefore, the specification of the confidence interval would not cause a problem. In the explanation related to the question, ChatGPT-3.5 mentioned at the beginning that knowing the alpha value is essential to test null and alternative hypotheses but proceeded by accepting the most commonly accepted value of 0.05. Therefore, adding the lack of a confidence interval value as attachments should not affect ChatGPT-3.5's answer as incorrect. ChatGPT-3.5 has given two different answers to this question: D) 0.059 and E) 0.04. However, the correct answer had to be only 0.059, but ChatGPT-3.5 has also pointed out 0.04 as the correct answer, where it is even less than 0.05. It seems that ChatGPT-3.5's mistake in this question was to compare incorrectly, in particular decimal numbers. However, as a different question, when asked "Is 0.05 smaller than 0.044?", ChatGPT answered that 0.044 is smaller than 0.05.

Both our study and other studies in the literature show that the answers provided by ChatGPT-3.5 regarding mathematical operations can be misleading. However, the performance of ChatGPT-4 was better than ChatGPT- 3.5, at least, based on the results of the biostatistics exam, although ChatGPT-4 still needs further training, we can say that it has reached a sufficient level of AI learning. Due to its ability to make comments related to biostatistics, select appropriate tests, and have a good command of basic biostatistics concepts, it can be beneficial to use ChatGPT as an education assistant in biostatistics education. Providing comprehensive and understandable explanations to questions can facilitate students' understanding of questions that they are confused about in relation to biostatistics.

It should not be forgotten that the main components of ChatGPT are the actor and critic models, which are trained using reinforcement learning with human feedback (RLHF). Therefore, even though ChatGPT is trained on large data, there is often the possibility of errors or oversights during the training process, and the training data itself may contain inaccurate information. In addition, in terms of educators, the fact that students can easily access ready-made information may harm their problem-solving and thinking skills. Relying solely on language models for homework, exams, or research can result in both unprogressive and uniform responses and ethical violations. However, another important point to remember is that ChatGPT, as an artificial intelligence language model, cannot access real-time information and its database is only up to date until a certain point. Therefore, it cannot provide direct information about current events or data. Instead, it can help us with information from the past and guide us on how to access up-to-date information.

The most significant limitation of this study is that ChatGPT receives regular updates, and its capabilities as a natural language processing model may vary across different versions. Since ChatGPT is a software that continuously updates and expands its database with each update, future versions of ChatGPT will be at a more advanced level than its current state. Therefore, the achievements of advanced versions of ChatGPT should also be considered in future studies. Additionally, the usability of natural language processing technologies other than ChatGPT in biostatistics education should be evaluated. The accessibility of different technologies by students can also enhance the efficiency of utilizing these technologies.

In conclusion, as it can be understood from the correct and comprehensive answers given to the questions, especially with the explanations created by ChatGPT-4, language models are promising in the field of education as students can obtain more detailed information and reveal their own experiences. The use of this technology in combination with traditional biostatistics teaching methods can provide advantages for both the educator and the student. Based on this, it is thought that ChatGPT has the possibility of being used as a training assistant in the field of biostatistics.

REFERENCES

Bhat YA, Saeed G, Sahel SG, Almesned A, Alqwaee A, Al-Akhfash A. 2022. Evaluation of Basic Statistical Knowledge Among Medical Residents Published Article. Cardiology & Vascular Research.

Brearley AM, Rott KW, Le LJ. 2023. A Biostatistical Literacy Course: Teaching Medical and Public Health Professionals to Read and Interpret Statistics in the Published Literature. *Journal of Statistics and Data Science Education*.

Celik Y. 2019. The Importance of Biostatistical Methods in the “Evidence-Based Medicine”. International Journal of Basic and Clinical Studies (IJBCS). 8(1):1-7.

Chiang CL, Zelen M. 1985. What Is Biostatistics?. Biometrics. 41(3):771.

Choi JH, Hickman KE, Monahan A, Schwarcz DB. ChatGPT Goes to Law School. 2023. Minnesota Legal Studies Research Paper No. 23-03. [accessed 2023 March 26]. http://dx.doi.org/10.2139/ssrn.4335905.

Couture F, Nguyen DD, Bhojani N, Lee JY, Richard PO. 2020. Knowledge and confidence level of Canadian urology residents toward biostatistics: A national survey. *Canadian Urological Association Journal*. 14(10).

Frieder S, Pinchetti L, Griffiths RR, Salvatori T, Lukasiewicz T, Petersen PC, Chevalier, A, Berner J. 2023. Mathematical Capabilities of ChatGPT (Version 1). arXiv:2301.13867 [accessed 2023 March 26]

Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. 2023. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education*, 9:e45312.

GPT-4 is OpenAI’s most advanced system, producing safer and more useful responses. 2023. California: OpenAI; [accessed 2023 March 26]. https://openai.com/product/gpt-4.

GPT-4. 2023. California: OpenAI; [Accessed 2023 March 26]. https://openai.com/research/gpt-4.

Gruzieva TS, Stuchynska NV, Inshakova HV. 2020. Research on the effectiveness of teaching biostatistics of future physicians. *Wiadomości Lekarskie*. 73(10):2227–2232.

Hanif A, Ajmal T. 2011. Statistical Errors in Medical Journals (A Critical Appraisal). Annals. 17(2):178-182.

Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, Weber T, Wesp P, Sabel B, Ricke J, Ingrisch M. 2022. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports (Version 1). arXiv.2212.14882. [accessed 2023 March 26]

KEYPS: Kurumsal Egitim Yonetim ve Planlama Sistemi. 2023. Ankara: KEYPS; [accessed 2023 March 26]. www.keyps.com.tr/.

Khan RA, Jawaid M, Khan AR, Sajjad M. 2023. ChatGPT - Reshaping medical education and clinical management. *Pakistan Journal of Medical Sciences*, 39(2).

Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health,* 2(2):e0000198.

Kurian N, Cherian JM, Sudharson NA, Varghese KG, Wadhwa S. 2023. AI is now everywhere. *British Dental Journal*, 234(2): 72–72.

Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. 2023. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digital Health. 2(2):e0000205.

Msaouel P, Kappos T, Tasoulis A, Apostolopoulos AP, Lekkas I, Tripodaki ES, Keramaris NC. 2014. Assessment of cognitive biases and biostatistics knowledge of medical residents: a multicenter, cross-sectional questionnaire study. *Medical Education Online*. 19(1):23646.

Singh JP, Neupane S, Mehta RK, Deo GP. 2022. Assessing undergraduate students' knowledge regarding application of biostatistics in research at medical college. *Journal of Chitwan Medical College*. 12(2):3–5.

Taecharungroj V. 2023. "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. Big Data and Cognitive Computing, 7(1):35

Talan, T. & Kalınkara, Y. (2023). The Role of Artificial Intelligence in Higher Education: ChatGPT Assessment for Anatomy Course. *Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi*, 7(1), 33-40. DOI: 10.33461/uybisbbd.1244777

Tomak L, Civanbay H. 2022. Evaluation of biostatistics knowledge and skills of medical faculty students. *Journal of Experimental and Clinical Medicine*. 19(3):620–627.

Vera-Ponce VJ, Torres-Malca JR, La Cruz-Vargas JAD, Zuzunaga Montoya FE, Chavez P H, Talavera-Ramirez JE, Cruz-Ausejo L. 2022. Analysis of Statistical Knowledge of Peruvian Medical Students: A Cross-Sectional Analytical Study Based on a Survey. *International Journal of Statistics in Medical Research*. 11:59–65.

Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, Fan Q, Wu S, Hu W, Li X. 2023. ChatGPT Performs on the Chinese National Medical Licensing Examination.