HAYDARPAŞA NUMUNE MEDICAL JOURNAL

DOI: 10.14744/hnhj.2024.92979
Haydarpasa Numune Med J 2025;65(1):19–24

hnhtipdergisi.com

ORIGINAL ARTICLE

Assessing the Competence of the ChatGPT-3.5 Artificial Intelligence System in Executing the ACLS Protocol of the AHA 2020

İbrahim Altundağ, Sinem Doğruyol, Burcu Genç Yavuz, Kaan Yusufoglu, Mustafa Ahmet Afacan, Şahin Çolak

Department of Emergency Medicine, University of Health Sciences Türkiye, Haydarpaşa Numune Training and Research Hospital, Istanbul, Türkiye

Abstract

Introduction: Artificial intelligence (AI) has become the focus of recent studies, particularly due to its potential to reduce human labor and time loss. The most significant contribution of AI applications in the medical field is expected to be enhancing clinicians' efficiency, reducing costs, and improving public health. This study aims to assess the proficiency of ChatGPT-3.5, one of the most advanced AI applications available today, in its knowledge of current information based on the American Heart Association (AHA) 2020 guidelines.

Methods: An 80-question quiz in a question-and-answer format, covering the current AHA 2020 application steps, was prepared and administered to ChatGPT-3.5 in both English (ChatGPT-3.5 English) and Turkish (ChatGPT-3.5 Turkish). The questions were originally prepared in Turkish for emergency medicine specialists.

Results: We found a similar success rate of over 80% in all questions posed to ChatGPT-3.5 and two independent emergency medicine specialists with at least five years of experience who did not know each other. ChatGPT-3.5 achieved a 100% success rate in all questions related to the General Overview of the Current AHA Guidelines, Airway Management, and Ventilation chapters in English.

Discussion and Conclusion: Our study indicates that ChatGPT-3.5 provides responses that are as accurate and up-to-date as those given by experienced emergency specialists regarding the AHA 2020 Advanced Cardiac Life Support Guidelines. With future updated versions of ChatGPT, instant access to accurate and current information based on textbooks and guidelines will be increasingly feasible.

Keywords: Artificial intelligence; AI chatbot; clinical decision support; generative pretrained transformer; guidelines.

One of the primary goals of developing technology and computer systems is to reduce dependence on human labor and create autonomous systems. Instead of systems that merely execute predetermined commands, the focus is on developing artificial intelligence (AI) systems capable of autonomously responding to changing conditions. By utilizing intelligent algorithms and iterative processes, AI mimics human intelligence and enhances its performance by continuously updating the vast amount of information it gathers.

Today, although not yet widely used for clinical decision support (CDS), one of the most notable examples of

Correspondence: İbrahim Altundağ, M.D. Department of Emergency Medicine, University of Health Sciences Türkiye, Haydarpaşa Numune Training and Research Hospital, Istanbul, Türkiye
Phone: +90 544 890 44 40 **E-mail:** dr.ibrahimaltundag@gmail.com
Submitted Date: 23.08.2024 **Revised Date:** 23.08.2024 **Accepted Date:** 24.09.2024
Haydarpaşa Numune Medical Journal

natural language processing (NLP) and CDS models is the Generative Pretrained Transformer (GPT) model.[1] The popularity of AI has recently surged in mainstream media and academic literature, particularly with the emergence of Generative Pretrained Transformer-3 (GPT-3), a language model capable of generating human-like text.[2]

Since the introduction of advanced AI applications like ChatGPT, its most prominent feature in the CDS field has been its use for case-based diagnosis and differential diagnosis purposes.[3,4] However, another crucial function of ChatGPT is its ability to instantly provide users with accurate and relevant information on specific topics. Despite the vast amount of medical information and resources available, accessing precise and reliable information remains a challenge. While the widespread use of the internet has facilitated information retrieval, clinicians may still struggle to access specific details efficiently. The ability to obtain accurate information quickly is especially critical for clinicians working in high-pressure environments such as emergency departments. AI-powered querying via ChatGPT, which can be regarded as a form of consultation or advisory support, is expected to be one of the key applications of AI in emergency medicine, allowing clinicians to access up-to-date and accurate information within seconds.

In this study, we aimed to assess the proficiency of ChatGPT-3.5, one of the most advanced AI applications available today, in terms of its knowledge of current information based on the American Heart Association (AHA) 2020 Advanced Cardiac Life Support (ACLS) guidelines. To achieve this, a quiz was prepared in both Turkish and English and presented to ChatGPT-3.5. The same quiz was also administered to emergency medicine specialists to compare ChatGPT-3.5's success rates in Turkish and English. Additionally, considering that clinicians may use ChatGPT-3.5 for queries in their native language, we aimed to evaluate its accuracy in both languages.

Materials and Methods

Our study was conducted at the Emergency Medicine Department of Haydarpaşa Numune Training and Research Hospital. Ethics committee approval was obtained from the Health Sciences University Haydarpaşa Numune Training and Research Hospital Clinical Research Ethics Committee (HNEAH-KAEK 2023/KK/140). This study was performed in accordance with the principles of the Declaration of Helsinki. Informed consent was obtained from the emergency medicine specialists enrolled in the study.

Study Design

A quiz consisting of 80 questions, divided into 8 different chapters, was designed to assess current practices regarding the American Heart Association (AHA) 2020 Advanced Cardiac Life Support (ACLS) application steps.[5] The prepared quiz was administered to both ChatGPT-3.5 and two emergency medicine specialists with a minimum of five years of experience who were not informed about the study protocol. The quiz results were evaluated by a blinded researcher.

In the study, the accuracy of ChatGPT-3.5's answers was first assessed with reference to the AHA 2020 ACLS guidelines. Additionally, the inter-rater agreement of the quiz items was analyzed. Subsequently, the accuracy rates of ChatGPT-3.5 and the emergency medicine specialists' answers were compared (Fig. 1).

The accuracy rates were evaluated both at the individual question level and chapter level and were reported as a "success percentage." Bloom's 80% cut-off value was used to determine the success of ChatGPT-3.5, in accordance with the literature.[6]

Quiz Components

An 80-question quiz was created in a question-and-answer format, covering the current ACLS application steps. These steps were structured into eight chapters, addressing key areas of the AHA 2020 ACLS guidelines:

1. General information on current ACLS guidelines (5 questions)
2. Basic-advanced airway management (7 questions)
3. High-quality cardiopulmonary resuscitation (CPR) (11 questions)
4. Ventilation (8 questions)
5. Defibrillation (12 questions)

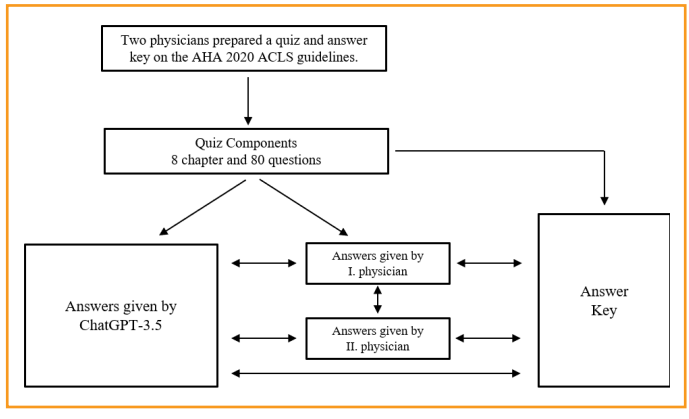

Figure 1. Flow diagram that illustrates the design of the study.

6. Medications (22 questions)
7. Vascular access and return of spontaneous circulation (ROSC) (6 questions)
8. Class of recommendation (9 questions)

In total, the quiz consisted of 80 questions aimed at evaluating the AI model's and emergency medicine specialists' knowledge of ACLS principles.

Obtaining Data

The quiz was administered using the ChatGPT-3.5 March 2023 version. It was applied to the AI model in both English (ChatGPT-3.5 English) and Turkish (ChatGPT-3.5 Turkish), whereas for emergency medicine specialists, the questions were prepared only in Turkish.

The quiz questions were designed to have definitive, unchanging, and non-interpretable answers. They were presented to ChatGPT-3.5 in plain text format, and the responses were evaluated for compliance with the AHA 2020 ACLS guidelines (Fig. 2).

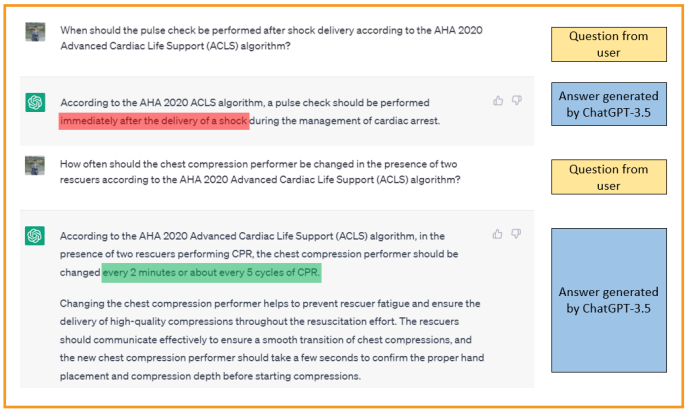

Figure 2. Scheme of the ChatGPT-3.5 usage. Example questions with wrong answers (at the top) and correct answers (at the bottom).

Outcomes

The primary aim of this study was to assess the accuracy of ChatGPT-3.5 in answering questions related to the current ACLS guidelines. The secondary aim was to compare the knowledge level of experienced emergency medicine specialists with that of ChatGPT-3.5 regarding ACLS. The accuracy of responses to ACLS-related questions was evaluated based on the *Adult Basic and Advanced Life Support: 2020 AHA Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care*.[5]

Statistical Analysis

In our study, categorical data were expressed as numbers (n) and percentages (%). The chi-square test was used for data comparisons. The agreement between specialists and ChatGPT-3.5 for categorical data was evaluated using inter-rater agreement analysis. Inter-rater agreement was expressed with the Kappa value and 95% confidence interval (CI), with the significance level set at $p<0.05$.

Data analysis was performed using MedCalc Version 20.218 (MedCalc Software Ltd, Ostend, Belgium) and IBM SPSS Version 20 (IBM Corp, Armonk, NY).

Results

The success rates for the entire quiz were as follows: I. specialist: 81.3% (65/80); II. specialist: 87.5% (70/80); ChatGPT-3.5 in Turkish: 81.3% (65/80); ChatGPT-3.5 in English: 86.3% (69/80).

There was no statistically significant difference in success rates between the I. specialist and II. specialist (p=0.067). Similarly, there was no statistically significant difference in success rates between the I. specialist and ChatGPT-3.5 in Turkish (p=0.386) or between the II. specialist and ChatGPT-3.5 in Turkish (p=0.332).

The distribution of correct answer rates and success rates by section is shown in Table 1. Additionally, in Table 1, the

Table 1. Comparison of success percentages of ChatGPT-3.5 and specialists according to Quiz sections

Chapters (questions)	I. Specialist	II. Specialist	ChatGPT-3.5 Turkish	ChatGPT-3.5 English	p*
A – General Overview for Current AHA Guideline (n=5)	5/5 (%100)	5/5 (%100)	5/5 (%100)	5/5 (%100)	-
B – Airway Management (n=7)	5/7 (%71.4)	7/7 (%100)	7/7 (%100)	7/7 (%100)	0.110
C – High Quality CPR (n=11)	8/11 (%72.7)	10/11 (%90.9)	11/11 (%100)	10/11 (%90.9)	0.137
D – Ventilation (n=8)	7/8 (%87.5)	8/8 (%100)	5/8 (%62.5)	8/8 (%100)	0.122
E – Defibrillation (n=12)	9/12 (%75)	10/12 (%83.3)	10/12 (%83.3)	9/12 (%75)	0.837
F – Medications (n=22)	20/22 (%90.9)	18/22 (%81.8)	16/22 (%72.7)	19/22 (%86.4)	0.295
G – Vascular Access and ROSC (n=6)	5/6 (%83.3)	4/6 (%66.6)	6/6 (%100)	5/6 (%83.3)	0.301
H – Class (Strength) of Recommendation (n=9)	6/9 (%66.6)	8/9 (%88.8)	5/9 (%55.5)	6/9 (%66.6)	0.288

AHA: American Heart Association; ROSC: Return of spontaneous circulation; CPR: Cardiopulmonary resuscitation. The ratio of correct answers to all answers is given in parentheses as a percentage. *P values obtained by comparing the success rates of ChatGPT-3.5 Turkish with the success rates of specialists.

Table 2. Inter-rater aggrement Kappa values

Agreement	Kappa Values	95% Confidence Interval
I.specialist – II.specialist	0.20	-0.06-0.46
I.specialist – ChatGPT-3.5 Turkish	0.09	-0.14-0.33
II.specialist – ChatGPT-3.5 Turkish	0.10	-0.13-0.35
ChatGPT-3.5 Turkish- ChatGPT-3.5 English	0.27	0.00-0.53

success rates of the specialists and ChatGPT-3.5 in Turkish were compared chapter by chapter, and no statistically significant difference was found between the two groups.

The Kappa values, obtained from the comparison of the correct/incorrect answers in terms of inter-rater agreement, are presented in Table 2. The only comparison that showed fair agreement was between ChatGPT-3.5's responses to the Turkish and English versions of the quiz (Kappa=0.27, p=0.015).

There were four questions for which both emergency medicine specialists provided incorrect answers. ChatGPT-3.5 in Turkish also failed to provide the correct answer to one of these questions (p=0.742). This question was:

"What is the maximum initial dose for biphasic defibrillators according to the AHA 2020 ACLS algorithm?"

However, it was observed that ChatGPT-3.5 provided the correct answer to this question in the English version of the quiz.

The number of questions that both specialists answered correctly was 59, and when analyzed by chapter, the highest common correct answer rate was 5 out of 5 (100%) in the 'A - General Overview for Current AHA Guidelines' chapter. The chapter with the second-highest common correct answer rate was 'D - Ventilation' with 87.5%.

ChatGPT-3.5 in Turkish provided the correct answer to 50 out of these 59 questions that both specialists answered correctly (p=0.179).

The test results of ChatGPT-3.5 and the specialists are uploaded as supplemental materials.

Discussion

To the best of our knowledge, this is the first study to evaluate ChatGPT-3.5's knowledge of the AHA 2020 ACLS guidelines. We observed a similar success rate of over 80% in all questions posed to ChatGPT-3.5 and two emergency medicine specialists working independently in different hospitals without prior knowledge of each other.

ChatGPT-3.5 answered all questions related to the General Overview for Current AHA Guidelines, Airway Management, and Ventilation chapters with a 100% success rate in English, while it achieved a 100% success rate in Turkish for the General Overview for Current AHA Guidelines, Airway Management, High-Quality CPR, and Vascular Access and ROSC chapters. Based on our findings, ChatGPT-3.5 provides highly accurate and up-to-date answers to questions about current ACLS practices. Additionally, its success rates were comparable to those of experienced emergency medicine specialists.

One of the most significant advantages of AI is its potential to reduce human labor and time loss. The promise of AI applications in healthcare lies in improving efficiency for clinicians, reducing costs, and enhancing public health outcomes. The key prerequisites for the successful implementation of AI in medicine are accessibility, standardization, quality, and the availability of representative data.[7]

Several studies have investigated ChatGPT's applications, including triage and differential diagnosis, drug interaction queries, AI-generated article writing, and medical problem-solving.[8-10] Most studies emphasize the clinical decision support (CDS) aspect of ChatGPT, where it makes inferences about variable, hypothetical cases, and the accuracy of these inferences is evaluated. However, studies like ours, which focus on evaluating ChatGPT as a reference source rather than for scenario-based reasoning, contribute to the literature by demonstrating an alternative application of AI in medicine.

The language in which ChatGPT is queried plays a crucial role when using it as a reference source. Since English is the primary language of medical literature, queries made in English generally yield more accurate responses. In our study, the difference in accuracy rates between English (86.3%) and Turkish (81.3%) supports the necessity of using English for optimal performance.

ChatGPT's reliance on plain text sources may have contributed to its poorer performance in the "Class of Recommendation" section, which was the least successful section in the quiz for ChatGPT-3.5 (66.6% accuracy in English and 55% accuracy in Turkish). Interestingly, this was also one of the lowest-performing sections for the specialists (66.6% and 88.8% accuracy, respectively). One potential explanation is that guideline recommendations are typically presented in tables, whereas ChatGPT primarily processes and presents information in plain text rather than structured formats like tables or figures. This

suggests that when performing medical literature queries through ChatGPT-3.5, users should verify the accuracy of information contained in tables, as ChatGPT may provide incorrect interpretations.

Another critical finding of our study is that ChatGPT-3.5 provided both correct and incorrect answers, which poses a risk of misleading clinicians and users. In our study, ChatGPT-3.5 answered 69 out of 80 questions correctly in English but provided incorrect answers to 11 questions. This deviation highlights the limitations of using ChatGPT as a definitive reference source. However, given that ChatGPT-3.5's accuracy was comparable to that of experienced specialists, and with future advancements in algorithms and enriched databases, the potential reduction in incorrect responses may enhance the feasibility of using ChatGPT as a reliable reference source in the coming years.

Limitations

One of the main limitations of our study is that ChatGPT-3.5 has not received clinical approval for obtaining healthcare information. Although our study demonstrated successful results with ChatGPT-3.5, it should be noted that AI applications, including ChatGPT, must be used with appropriate methodologies, and that ChatGPT is still under development, meaning its responses may be incorrect. A significant limitation of ChatGPT is that its incorrect answers can lead to misguidance and potential medical errors.[11]

Another important limitation of our study is the language issue. Different results may be obtained when queries are performed in the native language compared to English. Given that the medical literature is predominantly written in English, we believe that queries should be conducted in English, and medical terms should be searched using the standard forms found in the literature.

Additionally, our study did not impose a time limit for ChatGPT-3.5 or the emergency medicine specialists when answering each quiz question. Incorporating response time data into future studies may provide an opportunity to compare the efficiency of AI and human specialists in retrieving and processing information.

Conclusion

Our study demonstrated that ChatGPT-3.5 possesses a level of accurate and up-to-date knowledge comparable to that of an experienced emergency medicine specialist regarding the AHA 2020 Advanced Cardiac Life Support Guidelines. With the advancement of algorithms and the development of new versions, ChatGPT's mastery of current medical information can be further improved, reducing the number of incorrect responses.

Querying current guideline information through ChatGPT has the potential to serve as a readily accessible consultant function for emergency physicians. We believe that our study highlights the possibility of using ChatGPT as a portal for instant access to accurate and up-to-date information derived from textbooks and guidelines in the coming years.

Declaration: This article was previously published as a preprint on Research Square: https://doi.org/10.21203/rs.3.rs-3035900/v1.

Ethics Committee Approval: The study was approved by Health Sciences University Haydarpaşa Numune Training and Research Hospital Clinical Research Ethics Committee (No: HNEAH-KAEK 2023/KK/140, Date: 14/08/2023)

Peer-review: Externally peer-reviewed.

Use of AI for Writing Assistance: Not declared.

Authorship Contributions: Concept – S.D., K.Y., M.A.A.; Design – I.A., S.D., B.G.Y., K.Y., M.A.A., S.Ç.; Supervision – S.Ç.; Data collection &/or processing – I.A., S.D., B.G.Y., S.Ç.; Analysis and/or interpretation – I.A., S.D., K.Y., M.A.A.; Literature search – I.A., S.D., B.G.Y.; Writing – I.A., S.D., B.G.Y.; Critical review – S.D., B.G.Y.

Conflict of Interest: The authors declare that there is no conflict of interest.

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

References

1. Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care 2019;8:2328–31. [CrossRef]
2. Nath S, Marie A, Ellershaw S, Korot E, Keane PA. New meaning for NLP: The trials and tribulations of natural language processing with GPT-3 in ophthalmology. Br J Ophthalmol 2022;106:889–92. [CrossRef]
3. Goodwin TR, Harabagiu SM. Medical question answering for clinical decision support. Proc ACM Int Conf Inf Knowl Manag 2016;2016:297–306. [CrossRef]
4. Xu G, Rong W, Wang Y, Ouyang Y, Xiong Z. External features enriched model for biomedical question answering. BMC Bioinformatics 2021;22:272. [CrossRef]
5. Panchal AR, Bartos JA, Cabañas JG, Donnino MW, Drennan IR, Hirsch KG, et al. Part 3: Adult basic and advanced life support: 2020 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation 2020;142:S366–468. [CrossRef]
6. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Handbook 1: Cognitive domain. New York: David McKay; 1956.
7. Matheny ME, Whicher D, Thadaney Israni S. Artificial

intelligence in health care: A report from the National Academy of Medicine. JAMA 2020;323:509–10. [CrossRef]
8. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. Cureus 2023;15:e36272. [CrossRef]
9. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. Biol Sport 2023;40:615–22. [CrossRef]
10. Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. Cureus 2023;15:e35237. [CrossRef]
11. King MR. The future of AI in medicine: A perspective from a chatbot. Ann Biomed Eng 2023;51:291–5. [CrossRef]