Evaluating the Utility of ChatGPT in Diagnosing and Managing Maxillofacial Trauma

Evan Rothchild, BE, Caroline Baker, BA,* Isabelle T. Smith, BS,[†] Neil Tanna, MD, MBA,[‡] and Joseph A. Ricci, MD[‡]*

Abstract: Maxillofacial trauma is a significant concern in emergency departments (EDs) due to its high prevalence and the complexity of its management. However, many ED physicians lack specialized training and confidence in handling these cases, leading to a high rate of facial trauma referrals and increased stress on consult services. Recent advancements in artificial intelligence, particularly in large language models such as ChatGPT, have shown potential in aiding clinical decision-making. This study specifically examines the efficacy of ChatGPT in diagnosing and managing maxillofacial trauma. Ten clinical vignettes describing common facial trauma scenarios were presented to a group of plastic surgery residents from a tertiary care center and to ChatGPT. The chatbot and residents were asked to provide their diagnosis, ED management, and definitive management for each scenario. Responses were scored by attending plastic surgeons who were blinded to the response source. The study included 13 resident and ChatGPT responses. The mean total scores were similar between residents and ChatGPT (23.23 versus 22.77, $P > 0.05$). ChatGPT outperformed residents in diagnostic accuracy (9.85 versus 8.54, $P < 0.001$) but underperformed in definitive management (8.35 versus 6.35, $P < 0.001$). There was no significant difference in ED management scores between ChatGPT and the residents. ChatGPT demonstrated high accuracy in diagnosing maxillofacial trauma. However, its ability to suggest appropriate ED management and definitive treatment plans was limited. These findings suggest that while ChatGPT may serve as a valuable diagnostic tool in ED settings, further advancements are necessary before it can reliably contribute to treatment planning in emergent maxillofacial clinical scenarios.

Key Words: Artificial intelligence, ChatGPT, maxillofacial trauma

(*J Craniofac Surg* 2025;36: 1237–1241)

Maxillofacial trauma accounts for nearly one-fifth of all trauma admissions, making it a prevalent concern in emergency departments (EDs).[1] In a study by Trivedy et al,[2] 72.7% of ED physicians from various hospitals reported encountering at least one major facial trauma in the prior 3-month period, with 39.4% having seen over 20 cases. The etiology and patient presentations are diverse, ranging from soft tissue deformities to complex nasoorbitoethmoid fractures.[1,3–5] Consequently, management strategies vary greatly, involving laceration repair, intubation, joint reduction, foreign body removal, and other interventions.[1,5] Yet, numerous studies report that ED physicians lack formal training and knowledge in the diagnosis and management of maxillofacial trauma.[2,6,7] Of senior ED physicians surveyed in the study by Trivedy et al,[2] 27% did not feel confident managing facial trauma independently. However, only 28% to 42% of ED physicians across studies report having 24-hour maxillofacial support onsite for immediate consultation.[2,8,9] Thus, interfacility transfer rates for specialist referral are substantial, with Ray and colleagues reporting that 59% of all facial trauma cases presenting to a level 1 trauma center had been referred from another institution.[1,4,5,10]

Transfers add ~3 hours to a patient's time to treatment, despite ED physicians suggesting an acceptable time to treatment of 1.75 total hours.[2,5] These transfers also place an undue burden on specialized trauma centers. One level 1 trauma center in a rural state took 79% of transfers for facial trauma from the level 4 trauma centers and community hospitals comprising 84% of the state's hospitals. Unnecessary transfers in this study were estimated to cost $389,000 to $771,000, while no formal treatment is necessary for 29% to 41% of patients after transfer.[4] These findings suggest an opportunity to reduce the delay in care, the burden on high-level trauma centers, and cost through additional resources for diagnosis and ED management.[4,5]

Since their modern-day inception, large language models (LLMs), such as ChatGPT, have demonstrated the potential to augment the response of medical personnel in various ED scenarios. Studies highlight possible support in diagnosis, generating treatment plans, and improving operational efficiency.[11–16] However, the literature remains mixed, with some studies demonstrating poor performance.[16–18] To date, no known studies have explored LLM performance for facial trauma scenarios specifically. Therefore, given the persistent burden of referrals for facial trauma and the deficits in training, knowledge, and confidence of ED physicians in this area, we aim to investigate the efficacy of ChatGPT in ED diagnosis and initial and definitive management plans for cases of maxillofacial trauma.

From the *Division of Plastic Surgery, Montefiore Medical Center/Albert Einstein College of Medicine, New York, NY; †Vanderbilt University, Nashville, TN; and ‡Division of Plastic Surgery, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Great Neck, NY.

Received August 26, 2024.
Accepted for publication November 8, 2024.

Address correspondence and reprint requests to Joseph A. Ricci, MD, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell 600 Northern Boulevard, Great Neck, NY 11021; E-mail: jricci3@northwell.edu

The authors report no conflicts of interest.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.jcraniofacialsurgery.com.

ISSN: 1536-3732
DOI: 10.1097/SCS.0000000000010931

METHODS

A set of 10 clinical vignettes was created in June 2024, describing common facial trauma presentations to the ED (Supplemental Material, Supplemental Digital Content 1, http://links.lww.com/SCS/H124). These vignettes characterized a broad range of the most common facial injuries, standardized by beginning with a patient presentation, followed by written descriptions of relevant clinical history, physical examination findings, and imaging results. Each vignette was presented to a cohort of plastic surgery residents from a tertiary care center's plastic and reconstructive surgery residency program. Responders were then asked to provide the diagnosis, initial next steps in the ED, and definitive management, along with the clinical reasoning behind these choices. The responses and the year of training for each resident were recorded. The vignettes were subsequently presented to ChatGPT with the request for the same 3 open-ended answers. Each vignette was presented to ChatGPT the same number of times as there were resident responses to ensure an equal number in each group. Each prompt was presented to ChatGPT in separate conversations, and memory between conversations was turned off to prevent recursive learning from the previous vignettes.

Following the collection of responses from both the residents and ChatGPT, the responses were scored by 2 attending plastic and reconstructive surgeons blinded to the response's source. A standardized binary scoring system was used in which each component of the response (diagnosis, ED management, and definitive management) was scored as 0 or 1 for inaccurate or accurate, respectively. The highest score possible was 3 for each vignette and 30 in total. The mean score from the surgeons determined the final score of the responses. An overall score of 0 was proposed to represent poor clinical decision-making in patient presentations for facial trauma, whereas a score of 30 was considered excellent. In addition, the responses by the residents and ChatGPT were qualitatively compared by the plastic surgeons, noting any recognized patterns or major flaws between the two.

ChatGPT 4o was utilized because it is free of charge, thereby more closely mimicking the experience the majority of users will have. Statistical analysis was conducted using R version 4.4.1, and figures were created using GraphPad Prism version 10.3.0. The mean overall scores and scores for each of the 3 response categories between ChatGPT and the residents were compared using Mann-Whitney *U* tests. The interrater reliability between the attending plastic surgeon scorers was determined through the Pearson correlation coefficient. Linear regressions were used to determine the correlation between training level and response quality.

RESULTS

There were a total of 13 resident responses. The interrater reliability between the two plastic surgeons scoring was 0.919 ($P < 0.001$). The mean total resident response score was 23.23 (SD = 3.74), with a mean diagnosis score of 8.54 (SD = 0.69), ED management score of 6.35 (SD = 2.59) and a definitive management score of 8.35 (SD = 1.39). Among these were 6 junior resident responses, with a mean total response score of 22.92 (SD = 3.68), a diagnosis score of 8.50 (SD = 0.63), an ED management score of 6.00 (SD = 3.17), and a definitive management score of 8.42 (SD = 1.43). In addition, there were 7 senior resident responses, with a mean total response score of 23.50 (SD = 4.06), a diagnosis score of 8.57 (SD = 0.79), an ED management score of 6.64 (SD = 2.19), and a definitive management score of 8.29 (SD = 1.47). There were no significant differences between junior and senior residents in total, diagnosis, ED management, or definitive management scores. Furthermore, linear regression models showed no correlation between resident post-graduate year and total, diagnosis, ED management, or definitive treatment scores (Supplemental Digital Content, Table 1, http://links.lww.com/SCS/H125; Figs. 1, 2).

The mean ChatGPT total score was 22.77 (SD = 3.00). The mean diagnosis score was 9.85 (SD = 0.38), the emergency room (ER) management score was 6.58 (SD = 1.75), and the definitive management score was 6.35 (SD = 1.42). The total response and ER management scores between ChatGPT and the residents were not significantly different. However, ChatGPT's diagnosis scores were significantly higher than the residents' diagnosis scores ($P < 0.001$), while the residents' definitive management scores were significantly higher than those of ChatGPT ($P < 0.001$; Figs. 3, 4).

DISCUSSION

Artificial intelligence (AI) has been utilized in health care to analyze medical images, predict patient outcomes, and accelerate drug development.[19–21] More recently, with the introduction of LLMs, research has explored how chatbots like ChatGPT may assist providers in patient diagnosis and treatment.[11–16] Given the persistent challenges faced by EDs across the United States, such as overcrowding, understaffing, and delays in care, these AI tools hold significant potential to alleviate the burden on both ED providers and specialist consult services.[16] Motivated by the high rates of facial trauma referrals and the existing deficits in training, knowledge, and confidence among ED physicians, this study specifically examined the efficacy of ChatGPT in diagnosing and managing maxillofacial trauma in the ED. The integration of LLMs for these patients could streamline workflows, improve patient outcomes, and reduce stress for patients, providers, and large hospital systems.

In our study, ChatGPT scored nearly perfectly in diagnosing mock patient presentations of maxillofacial trauma, outperforming resident physicians with specialized knowledge. With access to a boundless database of online information and advanced natural language processing, ChatGPT can integrate diverse clinical knowledge to conduct contextually relevant analyses and determine probable characterizations of maxillofacial trauma cases. While this study appears to be the first to analyze maxillofacial trauma diagnosis in this way, ChatGPT has shown success in diagnosing other clinical conditions, including oral pathologies, otolaryngology vignettes, U.S. Medical Licensing Examination questions, and early undifferentiated

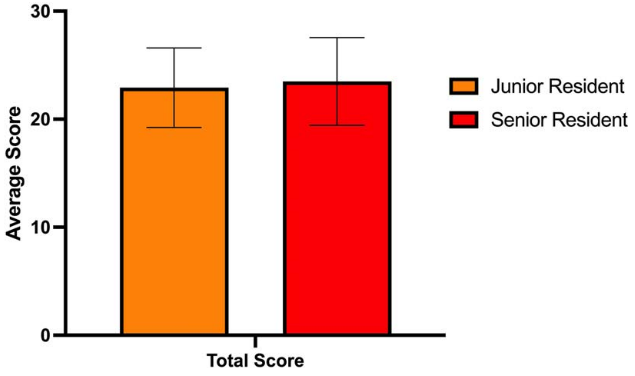

FIGURE 1. A graph that shows the average total scores between junior and senior residents.

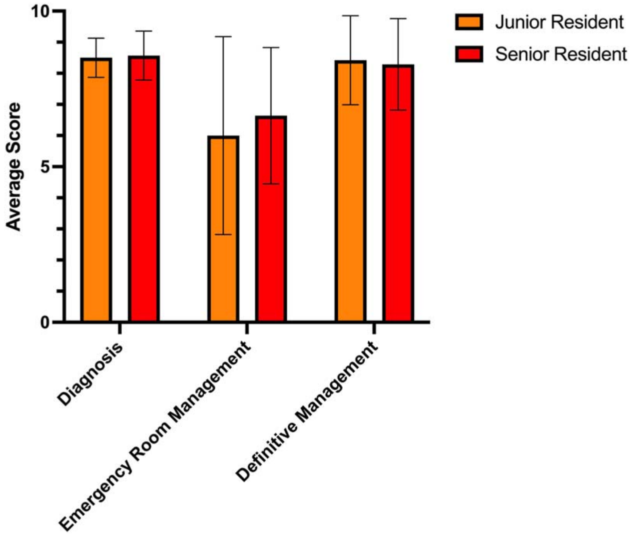

FIGURE 2. A graph that shows the average diagnosis, ER management, and definitive treatment scores between junior and senior residents. ER indicates emergency room.

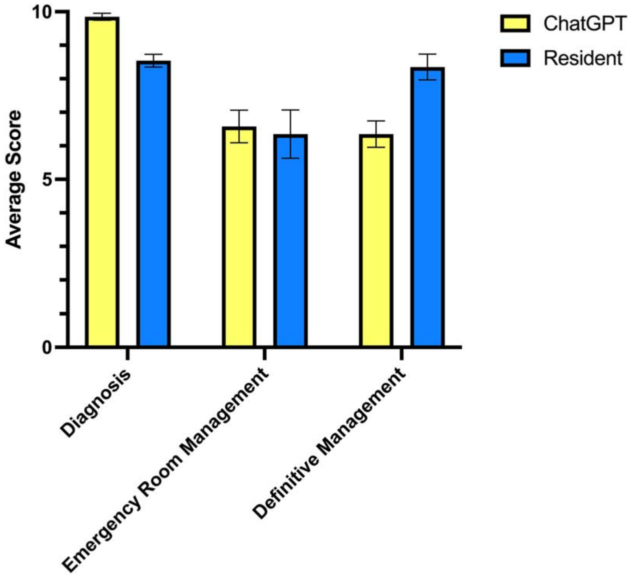

FIGURE 4. A graph that shows the average diagnosis, ER management, and definitive treatment scores between resident and ChatGPT respondents. ER indicates emergency room.

ED presentations.[15,22–24] Recent research also demonstrated a drastic improvement in medical diagnosis between iterations of ChatGPT, suggesting an upward trend in diagnostic capabilities.[23] Thus, our results add to a growing body of evidence suggesting that an LLM could serve as a supplementary tool for confirming or correcting diagnoses. While residents performed well in diagnosis with an accuracy of 85.4% in our study, this tool would provide an additional layer of reassurance, especially for ED physicians who are overloaded or untrained for these cases. Artificial intelligence tools can serve as diagnostic filters for ED physicians, helping to determine whether a consult is required or offering a reliable second opinion to reduce diagnostic errors. ED providers can use AI systems like ChatGPT as their initial reference before placing a consult, which is invaluable in trauma scenarios where time is a significant factor. This approach can streamline the diagnostic process, eliminate unnecessary consults, and relieve some burden on consult services or level 1 trauma centers. Although our results offer insight into ChatGPT's utility for diagnosing maxillofacial trauma cases, further research into rare and more complex presentations of facial trauma will be necessary to determine whether ChatGPT's clinical knowledge and analytical abilities continue to succeed. In addition, as this technology evolves, it will be essential to investigate whether recursive learning adds recency or other biases to predictions.

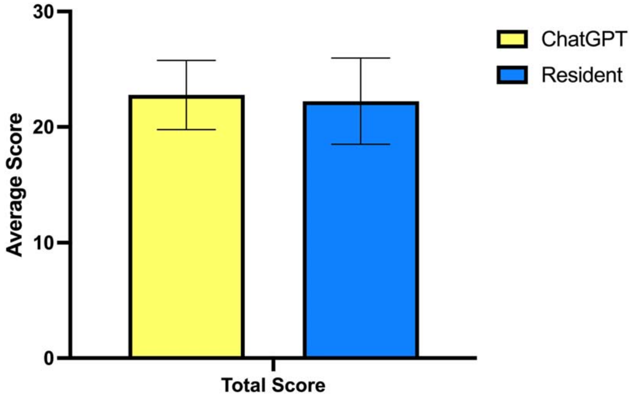

FIGURE 3. A graph that shows the average total scores between ChatGPT and residents.

ChatGPT lacked accuracy in determining appropriate ED management, although its performance was not significantly different from residents (65.8% and 63.5%, respectively). The program's deficiency in ED management may reflect its difficulty in understanding the nuances of patient stabilization and the prioritization of interventions in emergency protocols. There is currently a dearth of research on ChatGPT-predicted ED management plans, but limited research has also demonstrated that ChatGPT has low performance in ED triaging tasks.[17] Future research is necessary to describe this discrepancy between LLM results and best clinical ED management practices and to determine whether it exists for patient presentations beyond maxillofacial trauma.

Regarding definitive management, resident physicians outperformed ChatGPT with 83.5% and 63.2% accuracies, respectively. These results were surprising, and it is unclear why the suggested patient definitive management plans were not compliant with current clinical knowledge and guidelines. Investigations of ChatGPT's utility in other medical specialties have shown high accuracy and congruence between success in diagnosis and definitive management. For instance, in a study by Uranbey and colleagues, ChatGPT provided detailed treatment plans for classic nonemergent clinical presentations of oral pathologies, including further diagnostic requirements (consultations, imaging, and biopsies) and definitive treatments. These plans complied with those of medical professionals, scoring at least 80% in most cases.[22] Another study by Qu et al[15] demonstrated success in treatment planning for nonemergent otolaryngology-related patient presentations with a median score of 100% and no deterioration with increasing scenario complexity. Similarly, Rizwan and Sadiq[25] demonstrated high compliance between ChatGPT-determined diagnosis and treatment planning with current clinical knowledge and literature in cardiology. Our findings suggest that a current limitation of these systems is an inability to apply clinical information to suggest robust definitive treatment plans for maxillofacial trauma. Qualitative analysis of ChatGPT's responses revealed a trend of providing vague treatment plans with multiple possible solutions. The chatbot often suggested consulting the appropriate surgery service without offering a direct treatment plan. These findings are concerning because many maxillofacial trauma scenarios have distinct treatment algorithms that are crucial to follow. For example, in the case of an orbital fracture with entrapment, immediate surgery is recommended for every patient; however, ChatGPT failed to suggest this to the ER provider directly. Our findings indicate that ChatGPT is currently more effective as a diagnostic tool than assisting with emergent treatment strategies. However, as these models evolve, their ability to synthesize information and assist with treatment plans will likely improve significantly.

Additional interesting findings of our study were the general underperformance of residents, indicated by low total scores, and the lack of significant differences in treatment scores between junior and senior plastic surgery residents across diagnostic, ED, and definitive management scores. These findings were unexpected, as we predicted strong response scores from residents and an upward trend in the scores of plastic surgery residents as their knowledge improved with training. Several factors might explain these findings. One is the small sample size of 13 residents, which may not accurately reflect broader trends among plastic surgery residents. As we only surveyed plastic surgery residents and not otorhinolaryngology residents, this may also reflect either the specific training program's lack of emphasis on facial trauma or the absence of facial trauma calls for these plastic surgery residents. In addition, the clinical vignettes we created might not have accurately represented the complexity of real-life scenarios, limiting the senior residents' ability to demonstrate their complete expertise. Lastly, the survey format with limited instructions and minimums might have inherent limitations. The resident responses were shorter on average than the ChatGPT responses, suggesting that residents might not have been able to allocate enough time and thought to provide comprehensive answers.

This study likely provides a glimpse into a future where physicians may use AI models as clinical reference tools to enhance patient care. However, their implementation raises several essential considerations. First, the effectiveness of ChatGPT is directly correlated to the information the user provides. In real-world settings, the provider may input inaccurate or incomplete clinical information, making these systems susceptible to bias. This makes these tools user-dependent, highlighting the need for training programs to assist providers in using them effectively. Next, implementing chatbots raises liability questions. If the chatbot provides an incorrect diagnosis or treatment, is the provider who used the program liable, or are additional steps required to verify the AI's output? This is particularly relevant in maxillofacial trauma, where diagnosis and treatment are time-sensitive and can significantly influence patient outcomes. Lastly, their implementation raises concerns about overreliance on these tools. It is essential to stress that these tools are designed to supplement, not replace, the medical decision-making of trained medical personnel.

Limitations

Our study has several limitations. The most notable was the small sample size of the resident group assessed because the study was conducted at a single institution. Despite this critical limitation, we found statistically significant differences between the responses from the residents and ChatGPT. It will be crucial for future studies to use larger sample sizes to improve the reliability and generalizability of these findings. In addition, the responses were open-ended and without word count or keyword requirements. This lack of standardization may have impacted the resident responses, which were shorter on average than the thorough responses from ChatGPT. More extended responses from residents may have incidentally incorporated keywords necessary for higher scores. Also, future studies may compare the resident scores on these questions to validated assessments such as the in-service examination scores from the prior year to assess whether performance on these vignettes correlates with established measures of clinical knowledge. Another notable limitation is the style and quality of the vignettes we created. Despite several measures to ensure accurate, well-rounded clinical descriptions, the responses from both ChatGPT and the residents relied heavily on the information provided. It is possible that the results were more a reflection of the specific wording in the vignettes than an accurate reflection of the true capabilities and knowledge of ChatGPT and the residents. Furthermore, as a preliminary investigation, the vignettes used in this study distilled the information into a succinct paragraph with only relevant details. Scenarios with extraneous information that require the chatbot to identify pertinent positives or negatives may not yield equally robust results. Future studies that require ChatGPT to determine the important information would provide additional valuable insights into the clinical utility of these chatbots. Moreover, ChatGPT's training and accessible knowledge base may have impacted its analytical abilities. Our clinical vignettes were classic representations of different types of facial trauma, which are likely similar to cases on which ChatGPT has been trained. This may have enhanced the system's ability to diagnose accurately. More convoluted real-world scenarios in the hospital may not have such strong diagnostic results. Future IRB-approved studies should investigate ChatGPT's ability to accurately diagnose and develop treatment plans for real-world complex patient presentations. Finally, despite attempts to limit subjectivity, the attending surgeons' grading of resident and ChatGPT responses remains subjective, and sources of responses may be indicated by length and style of response. Notwithstanding these limitations, we demonstrate the ability of ChatGPT to serve as a diagnostic reference tool in managing cases of maxillofacial trauma and discuss the impact of AI chatbot implementation in the ED setting.

CONCLUSION

Overall, this study's findings suggest that ChatGPT can serve as an effective diagnostic reference for ER physicians managing maxillofacial trauma, optimizing the diagnostic process. However, current chatbot models struggle with creating specific ER and definitive treatment plans. As these tools continue to improve and become more integrated into medical practice, multiple considerations must be addressed to ensure they benefit patient care. While currently limited, these tools offer significant potential to assist ED physicians and alleviate the burden on maxillofacial consult services and level 1 trauma centers.

REFERENCES

1. Al-Hassani A, Ahmad K, El-Menyar A, et al. Prevalence and patterns of maxillofacial trauma: a retrospective descriptive study. *Eur J Trauma Emerg Surg* 2022;48:2513–2519
2. Trivedy C, Kodate N, Ross A, et al. The attitudes and awareness of emergency department (ED) physicians towards the management of common dentofacial emergencies. *Dent Traumatol* 2012;28:121–126
3. Trivedy C, Kodate N, Ross A, et al. The attitudes and awareness of emergency department (ED) physicians towards the management of common dentofacial emergencies. *Dent Traumatol.* 2020;28:121–3.

4. Ray A, Curti S, Pegues J, et al. Secondary overtriage of isolated facial trauma. *Am J Otolaryngol* 2021;42:103043
5. Pontell ME, Colazo JM, Drolet BC. Unnecessary interfacility transfers for craniomaxillofacial trauma. *Plast Reconstr Surg* 2020; 145:975e–983e
6. Patel KK, Driscoll P. Dental knowledge of accident and emergency senior house officers. *Emerg Med J* 2002;19:539–541
7. Addo ME, Parekh S, Moles DR, et al. Knowledge of dental trauma first aid (DTFA): the example of avulsed incisors in casualty departments and schools in London. *Br Dent J* 2007;202:E27
8. Whipple LA, Kelly T, Aliu O, et al. The crisis of deficiency in emergency coverage for hand and facial trauma: exploring the discrepancy between availability of elective and emergency surgical coverage. *Ann Plast Surg* 2017;79:354–358
9. Allonby-Neve CL, Okereke CD. Current management of facial wounds in UK accident and emergency departments. *Ann R Coll Surg Engl* 2006;88:144–150
10. Jose A, Nagori SA, Agarwal B, et al. Management of maxillofacial trauma in emergency: an update of challenges and controversies. *J Emerg Trauma Shock* 2016;9:73–80
11. Ayoub M, Ballout AA, Zayek RA, et al. Mind + machine: ChatGPT as a basic clinical decisions support tool. *Cureus* 2023;15:e43690
12. Sarraju A, Bruemmer D, Van Iterson E, et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* 2023;329:842–884
13. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine* 2023;183: 589–596
14. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLos Digit Health* 2023;2:e0000198
15. Qu RW, Qureshi U, Petersen G, et al. Diagnostic and management applications of chatgpt in structured otolaryngology clinical scenarios. *OTO Open* 2023;7:e67
16. Preiksaitis C, Ashenburg N, Bunney G, et al. The role of large language models in transforming emergency medicine: scoping review. *JMIR Med Inform* 2024;12:e53787
17. Sarbay İ, Berikol GB, Özturan İU. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): a preliminary, scenario-based cross-sectional study. *Turk J Emerg Med* 2023;23:156–161
18. Kim JH, Kim SK, Choi J, et al. Reliability of ChatGPT for performing triage task in the emergency department using the Korean Triage and Acuity Scale. *Digital health* 2024;10: 20552076241227132
19. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023;6:1169595
20. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. *Nat Med* 2022;28:31–38
21. Beam AL, Drazen JM, Kohane IS, et al. Artificial intelligence in medicine. *N Engl J Med* 2023;388:1220–1221
22. Uranbey Ö, Özbey F, Kaygısız Ö, et al. Assessing ChatGPT's diagnostic accuracy and therapeutic strategies in oral pathologies: a cross-sectional study. *Cureus* 2024;16:e58607
23. Shieh A, Tran B, He G, et al. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. *Sci Rep* 2024;14:9330
24. Berg HT, van Bakel B, van de Wouw L, et al. ChatGPT and generating a differential diagnosis early in an emergency department presentation. *Ann Emerg Med* 2024;83:83–86
25. Rizwan A, Sadiq T. The use of AI in diagnosing diseases and providing management plans: a consultation on cardiovascular disorders with ChatGPT. *Cureus* 2023;15:e43106

This publication is referenced:

The Biologic Basis for the Clinical Application of the Silicones

Full-text available Article August 1984 · Archives of Surgery

31 Reads · 101 Citations

The state of the art still stands up after more than 40 years