Generative artificial intelligence (AI) chatbots, like ChatGPT, have become more competent and prevalent, making their role in patient education more salient. A study aimed to compare the educational utility of six AI chatbots by quantifying the readability and quality of their answers to common patient questions about clavicle fracture management. Methods ChatGPT 4, ChatGPT 4o, Gemini 1.0, Gemini 1.5 Pro, Microsoft Copilot, and Perplexity were used with no prior training. Ten representative patient questions about clavicle fractures were posed to each model. The readability of AI responses was measured using Flesch-Kincaid Reading Grade Level, Gunning Fog, and Simple Measure of Gobbledygook (SMOG). Six orthopedists blindly graded the response quality of each model using the DISCERN criteria. Both metrics were analyzed via the Kruskal-Wallis test. Results No statistically significant difference was found among the readability of the six models. Microsoft Copilot (70.33±7.74) and Perplexity (71.83±7.57) demonstrated statistically significant higher DISCERN scores than ChatGPT 4 (56.67±7.15) and Gemini 1.5 Pro (51.00±8.94) with similar findings seen between Gemini 1.0 (68.00±6.42) and Gemini 1.5 Pro. The mean overall quality (question 16, DISCERN) of each model was rated at or above average (range, 3-4.4). Conclusion The findings suggest generative AI models have the capability to serve as supplementary patient education materials. With equal readability and overall high quality, Microsoft Copilot and Perplexity may be implicated as chatbots with the most educational utility regarding surgical intervention for clavicle fractures [1)].

Critical Review of the Study on Generative AI Chatbots in Patient Education for Clavicle Fractures

Introduction Generative AI chatbots are increasingly being considered for patient education, offering accessible and scalable information. This study sought to compare six AI chatbots (ChatGPT 4, ChatGPT 4o, Gemini 1.0, Gemini 1.5 Pro, Microsoft Copilot, and Perplexity) by evaluating the readability and quality of their responses to common patient questions about clavicle fracture management. The study employed objective readability metrics (Flesch-Kincaid, Gunning Fog, and SMOG) and the DISCERN criteria, rated by orthopedic surgeons, to assess the quality of responses.

While the study provides valuable insights, certain methodological limitations and broader implications must be critically examined.

Strengths of the Study 1. **Objective Comparison Across Multiple AI Models**

1. The inclusion of six major AI models allows for a broad comparative analysis.
2. Readability metrics provide standardized, quantifiable measures of text complexity.
3. The use of DISCERN, a validated tool for assessing health information quality, ensures a structured and expert-driven evaluation.

2. **Blinded Evaluation by Orthopedists**

1. The study mitigates bias by having independent orthopedic surgeons assess response quality.
2. This adds clinical relevance, ensuring that the answers align with real-world patient education needs.

3. **Use of Statistical Analysis**

1. The Kruskal-Wallis test is appropriate for comparing multiple independent groups when data may not follow a normal distribution.
2. The study effectively identifies statistically significant differences in response quality among the

models.

Limitations and Criticisms 1. **Limited Scope of Medical Topic (Clavicle Fractures Only)**

1. While clavicle fractures are a common orthopedic condition, the findings may not generalize to other medical topics.
2. AI chatbots may perform differently across specialties, requiring broader investigations across multiple health conditions.

2. **Lack of Contextual and Conversational Analysis**

1. Patient education is not solely about readability and factual accuracy.
2. Factors like empathetic language, conversational coherence, and adaptability to follow-up questions were not evaluated.
3. Some AI models (e.g., ChatGPT) may be designed for better conversational engagement rather than rigid factual responses.

3. **Potential Bias in DISCERN Evaluation**

1. Although DISCERN is a validated tool, its interpretation can still be subjective.
2. Different raters might weigh certain aspects differently, leading to variability in scoring.

4. **Exclusion of Prior Training or Optimization**

1. The study explicitly states that the AI models were used “with no prior training.”
2. In practice, chatbots can be fine-tuned for medical education, which could significantly alter their performance.
3. Evaluating untrained models may not reflect their real-world potential.

5. **Clinical Safety and Misinformation Risks Not Addressed**

1. The study focuses on readability and quality but does not assess misinformation risks or medical accuracy in depth.
2. Certain AI chatbots may generate outdated or incorrect information, which could be harmful in clinical contexts.
3. There is no mention of fact-checking or validation against established medical guidelines.

Implications and Future Research Directions 1. **Broader Medical Applications**

1. Future studies should examine AI chatbots across multiple medical domains, including chronic diseases, post-operative care, and emergency medicine.

2. **Patient-Centered Evaluation**

1. Instead of relying solely on expert evaluations, incorporating patient feedback would provide insights into real-world usability and trustworthiness.

3. **Conversational Adaptability and Emotional Intelligence**

1. Evaluating chatbots on their ability to engage in empathetic, adaptive dialogue would enhance understanding of their role in patient education.

4. **Longitudinal Studies on AI Integration in Clinical Settings**

1. Research should assess how AI chatbot recommendations influence patient decision-making and adherence to medical advice over time.

Conclusion The study provides a useful comparative analysis of AI chatbots for patient education on clavicle fractures, demonstrating that Microsoft Copilot and Perplexity performed best in response quality. However, its narrow focus, lack of conversational analysis, and exclusion of clinical safety considerations highlight the need for further research. Future studies should explore AI chatbots' broader medical applications, misinformation risks, and their integration into clinical practice.

1)
Giammanco PA, Collins CE, Zimmerman J, Kricfalusi M, Rice RC, Trumbo M, Carlson BA, Rajfer RA, Schneiderman BA, Elsissy JG. Evaluating the Quality and Readability of Information Provided by Generative Artificial Intelligence Chatbots on Clavicle Fracture Treatment Options. Cureus. 2025 Jan 9;17(1):e77200. doi: 10.7759/cureus.77200. PMID: 39925539; PMCID: PMC11806961.