Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Assessing AI in Various Elements of Enhanced Recovery After Surgery (ERAS)-Guided Ankle Fracture Treatment: A Comparative Analysis with Expert Agreement

Rui Wang[1], Xuanming Situ[1], Xu Sun[2], Jinchang Zhan[1], Xi Liu[3]

[1]Department of Orthopaedic, Zhongshan City Orthopaedic Hospital, Zhongshan, Guangdong Province, People's Republic of China; [2]Department of Orthopaedic Trauma, Beijing Jishuitan Hospital, Beijing, People's Republic of China; [3]Department of Sports, Sun Yat-sen Memorial Primary School, Zhongshan, Guangdong Province, People's Republic of China

Correspondence: Rui Wang, Department of Orthopaedic, Zhongshan City Orthopaedic Hospital, Zhongshan, Guangdong Province, People's Republic of China, Tel +86 18699690080, Fax +86 760 89907577, Email falcon2959@me.com

Objective: This study aimed to assess and compare the performance of ChatGPT and iFlytek Spark, two AI-powered large language models (LLMs), in generating clinical recommendations aligned with expert consensus on Enhanced Recovery After Surgery (ERAS)-guided ankle fracture treatment. This study aims to determine the applicability and reliability of AI in supporting ERAS protocols for optimized patient outcomes.

Methods: A qualitative comparative analysis was conducted using 35 structured clinical questions derived from the Expert Consensus on Optimizing Ankle Fracture Treatment Protocols under ERAS Principles. Questions covered preoperative preparation, intraoperative management, postoperative pain control and rehabilitation, and complication management. Responses from ChatGPT and iFlytek Spark were independently evaluated by two experienced trauma orthopedic specialists based on clinical relevance, consistency with expert consensus, and depth of reasoning.

Results: ChatGPT demonstrated higher alignment with expert consensus (29/35 questions, 82.9%), particularly in comprehensive perioperative recommendations, detailed medical rationales, and structured treatment plans. However, discrepancies were noted in intraoperative blood pressure management and preoperative antiemetic selection. iFlytek Spark aligned with expert consensus in 22/35 questions (62.9%), but responses were often more generalized, less clinically detailed, and occasionally inconsistent with best practices. Agreement between ChatGPT and iFlytek Spark was observed in 23/35 questions (65.7%), with ChatGPT generally exhibiting greater specificity, timeliness, and precision in its recommendations.

Conclusion: AI-powered LLMs, particularly ChatGPT, show promise in supporting clinical decision-making for ERAS-guided ankle fracture management. While ChatGPT provided more accurate and contextually relevant responses, inconsistencies with expert consensus highlight the need for further refinement, validation, and clinical integration. iFlytek Spark's lower conformity suggests potential differences in training data and underlying algorithms, underscoring the variability in AI-generated medical advice. To optimize AI's role in orthopedic care, future research should focus on enhancing AI alignment with medical guidelines, improving model transparency, and integrating physician oversight to ensure safe and effective clinical applications.

Keywords: artificial intelligence, AI, enhanced recovery after surgery, ERAS, ankle fracture, comparative analysis, medical decision-making, interdisciplinary collaboration, ChatGPT

Introduction

Ankle fractures are common traumatic injuries in clinical settings, with their incidence rising due to increased societal activities. Studies show a prevalence of 169 cases per 100,000 people annually.[1] The ankle, which is essential for weight-bearing, plays a crucial role in daily activities. Its anatomical and biomechanical properties make it susceptible to various fractures, ranging from simple non-displaced to complex intra-articular types. The primary treatment objectives are to

Received: 25 November 2024
Accepted: 6 March 2025
Published: 19 March 2025

restore joint structure and function, relieve pain, and improve the patient's quality of life.[2] However, the effectiveness of current treatments is being questioned. A US survey (2016–2019) reported an incidence rate of 14.1 per 10,000 person-years, with 3.3% of cases requiring surgery. Postoperative complications are significant, with about 10% of patients needing additional surgeries, and nearly a third experiencing residual pain.[3] Another study found major postoperative complications in 31.5% of cases, including residual pain, deep infections, malunion, post-traumatic osteoarthritis, implant breakage, complex regional pain syndrome, and joint contracture, with 21.7% requiring further surgical interventions.[4]

Advancements in medical technology and growing patient needs have led to evolving treatment strategies for ankle fractures. The Enhanced Recovery After Surgery (ERAS) concept, a modern approach, is increasingly applied in ankle fracture care. In 2019, Chinese medical experts, including those from Beijing Jishuitan Hospital, released the "Expert Consensus on Optimizing Ankle Fracture Treatment based on ERAS concept".[5] This concept emphasizes comprehensive perioperative strategies to reduce complications, shorten hospital stays, and speed up recovery, covering the preoperative, intraoperative, and postoperative phases. However, implementing these multidimensional management and diagnostic decisions is complex. Although expert consensus provides guidance, patient variability and resource limitations in primary healthcare institutions challenge the implementation of ERAS, raising the question of whether simpler, more reliable management methods are viable for these patients.

The rapid growth and evolution of large artificial intelligence (AI) models such as ChatGPT have significantly impacted various industries. The exploration of AI's role in orthopedic diagnosis and treatment is becoming increasingly important. AI's potential in the medical field is expanding, especially in diagnosing diseases, planning treatments, and managing patients.[6] For instance, AI algorithms can more accurately identify fracture types and severity, thereby improving clinical treatment recommendations.[7] FDA-approved AI detection tools are already being used for fracture detection.[8] AI's influence in orthopedics is evident in image recognition, risk prediction, patient payment models, and clinical decision-making.[9] Therefore, the aim of this study is to compare AI model recommendations with expert consensus under ERAS guidance in diagnosing and treating ankle fractures. It evaluates whether AI can align with expert consensus in managing these fractures and explores the potential of AI, particularly ChatGPT, in making decisions about treatment. As AI continues to play a significant role in healthcare, identifying areas for improvement and innovation is crucial for overcoming existing challenges. As Chow et al pointed out,

> In the ever-evolving AI-assisted healthcare conversations, identifying opportunities for improvement and advancement is crucial for addressing existing challenges and shaping the future trajectory of this dynamic field.

This highlights the significant importance of the present study.[10]

Research Methods

This research focused on the application of AI in ankle fracture treatment within the framework of ERAS principles and contrasts it with expert consensus via qualitative analysis. We scrutinized the replies of two AI models, ChatGPT and iFlytek Spark, to crucial queries derived from the "Expert Consensus on Optimizing Ankle Fracture Treatment Protocols under ERAS Principles".[5] The study was designed and completed according to the following workflow (Figure 1).

Selection of Questions

The research team precisely selected 29 essential questions from the "Expert Consensus on Optimizing Ankle Fracture Treatment Protocol Under ERAS Principles". These questions covered a wide array of topics, including patient admission, discharge, preoperative preparation, intraoperative management, postoperative pain control, rehabilitation activities, and extended follow-up. To maintain the integrity and comparability of the responses, 6 additional follow-up

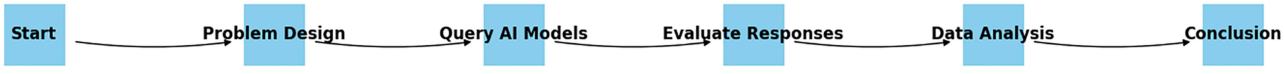

Figure 1 Workflow.

questions were strategically incorporated into the Q&A process. Due to space constraints, the complete list of 35 questions is provided in Appendix 1.

Choice and Use of AI Models

This study began by examining ChatGPT 4.0, recognized for its exceptional performance in various fields. Subsequently, GPT-4 was utilized to identify five advanced large language models, each containing over a hundred billion parameters, known for their effectiveness in Chinese language processing and applicability to medical research. To facilitate comparative analysis, iFlytek Spark was randomly chosen as a representative model. The research team then entered 35 Chinese questions into both AI models on a single day (December 3, 2023) and recorded their responses.

Method of Comparative Analysis

In December 2023, we accessed ChatGPT and iFlytek Spark through their official interfaces. Each of the 35 questions was inputted uniformly, without any prompt engineering or context adjustments. Two trauma orthopedic experts, each with over 10 years of experience, independently assessed the AI-generated responses. Evaluations were based on a binary scale: Consistent: The AI's response aligns closely with expert consensus, reflecting accurate clinical knowledge and recommendations. Inconsistent: The AI's response deviates significantly from expert consensus, potentially containing incorrect or misleading clinical information. Evaluators categorized each response accordingly and documented their rationale.

Example of AI Interaction

Below is a representative example of one question we focused on during the study, derived from questions 19–21 in the list: Example Question: "What is the appropriate intraoperative blood pressure control level for a patient undergoing ankle fracture surgery based on preoperative blood pressure?" The expert consensus recommends intraoperative blood pressure control to be maintained at 70–80% of the baseline blood pressure. ChatGPT suggested maintaining the blood pressure within 20% above or below the patient's normal range, whereas iFlytek Spark recommended a fixed range of SBP: 120–140mmHg and DBP: 80–90mmHg, We observed that both AI models provided recommendations that differed from the expert consensus. To further investigate the logic behind these responses, we posed follow-up questions to both models. The second question asked:

> Once the patient's pain is stable and based on vital signs, is it appropriate to maintain blood pressure at 70-80% of preoperative baseline for ankle fracture surgery?

ChatGPT disagreed, explaining that this was not the best practice, and reiterated that blood pressure should remain within 20% of the normal range. iFlytek Spark, in contrast, initially supported its previous recommendation but later revised its position to align with the Expert Consensus's approach, stating that maintaining blood pressure at 70–80% of baseline is a common and acceptable practice.

To validate ChatGPT's recommendation, we asked for supporting evidence, and it cited the American Society of Anesthesiologists' perioperative blood pressure management guidelines. We reviewed the guideline, which stated that intraoperative hypotension (a decrease in systolic blood pressure by 20% below baseline) might be associated with adverse outcomes, supporting ChatGPT's recommendation. On the other hand, when we asked iFlytek Spark for its evidence supporting the 70–80% baseline blood pressure range, It mentioned several studies, but aside from the ASA guidelines, no specific sources were provided for the other studies. However, after checking the ASA guidelines, we found no such recommendation. Interestingly, upon reviewing the ERAS Expert Consensus for ankle fracture perioperative management, we found that 97.1% of experts recommended maintaining blood pressure at 70–80% of baseline, citing a 2003 study on upper-limb tourniquet usage.

Data Analysis

We tallied the number of "consistent" responses for each AI model out of the 35 questions. Additionally, inter-rater agreement between the two experts was calculated to assess evaluation reliability. Initially, we designed a scoring system

ranging from 1 to 3 to evaluate the consistency of the responses generated by the two AI models: 1 for "inconsistent", 2 for "partially consistent", and 3 for "completely consistent". However, upon evaluation, it was interesting to note that neither of the two evaluators provided any ratings of "partially consistent". As a result, we decided to adopt a binary classification system for the final evaluation.

Results

The expert consensus on ERAS applied to orthopedic perioperative management is a comprehensive approach to perioperative care. The advantages of applying ERAS in orthopedic perioperative management have been confirmed by several studies. The management pathway encompasses various stages, including preoperative preparation (eg, nutrition, management of hypertension and diabetes), intraoperative management (eg, blood pressure control, use of tourniquets and catheters, surgical techniques), and postoperative rehabilitation. The 29 questions we designed, along with the 6 follow-up questions, cover multiple stages and dimensions of the ERAS management concept, enabling a comprehensive evaluation of AI's ability to apply these principles effectively.

During the question-answering process, the study formulated 29 core questions, supplemented by 6 additional questions (Q2, Q6, Q7, Q20, Q21, Q28) that required further clarification from the AI models (ChatGPT as AI1 and iFlytek Spark as AI2). Within this framework, two questions (Q1, Q2) were identified where both AI models provided consistent responses diverging from the expert consensus. Conversely, in nine instances (Q3, Q8, Q11, Q14, Q15, Q16, Q17, Q26, Q30), the expert consensus aligned with the responses of AI1 but not with those of AI2. For two questions (Q20, Q21), the expert consensus agreed with AI2 but not with AI1. Additionally, both AI models concurred on one question (Q33), although their responses contradicted the expert consensus. Notably, consensus was reached across all parties for 20 questions (Q4, Q5, Q6, Q7, Q9, Q10, Q12, Q13, Q18, Q22, Q23, Q24, Q25, Q27, Q28, Q29, Q31, Q32, Q34, Q35), while one question (Q19) elicited divergent views from each.

Examination of all 35 questions revealed that AI1's responses aligned with the expert consensus in 29 instances, while AI2's responses matched in 22 cases. Additionally, both AI models displayed consistent responses in 23 questions. These findings are visually depicted in a bar chart, illustrating the comparative analysis across various groups (Figure 2).

ChatGPT consistently matched the expert consensus in most questions, especially showcasing strong agreement in detailed quantitative measures. This was evident in parameters like preoperative blood glucose regulation and strategies

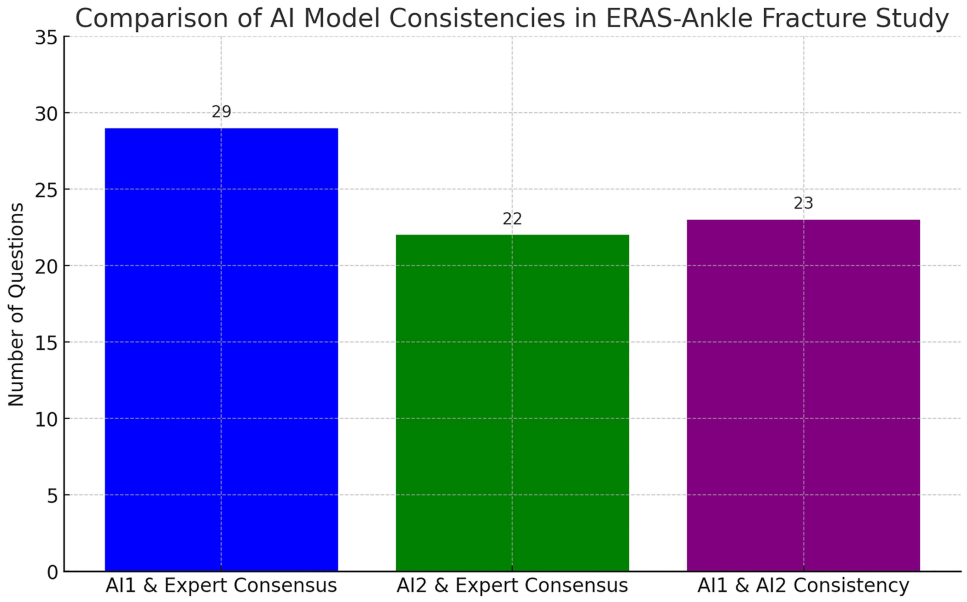

Figure 2 AI1 demonstrated alignment with the expert consensus in 29 questions, while AI2 matched the expert consensus in 22 questions. Additionally, there were 23 questions where both AI1 and AI2 exhibited mutual consistency with the expert consensus (Figure 2 was created by ChatGPT).

for assessing malnutrition. Additionally, when presented with well-defined follow-up queries, the AI proficiently identified the data sources supporting its decisions and participated in relevant discussions.

However, certain disparities became apparent. For example, in a question regarding intraoperative blood pressure control, the expert consensus advised maintaining blood pressure at 70–80% of baseline, while ChatGPT recommended a ± 20% range from normal blood pressure. Highlighting this inconsistency, ChatGPT referenced a 2021 anesthesiology article discussing the risks associated with lowering blood pressure below 20% of baseline.[11] Consequently, it challenged the expert consensus. Conversely, iFlytek Spark's response, which suggested maintaining blood pressure within the standard 120–140mmHg systolic and 80–90mmHg diastolic range, was considered less adaptable to the variability of patients in clinical settings.

Inconsistencies were also observed in preoperative acute pain management. While ChatGPT concurred with the expert consensus regarding oral acetaminophen or NSAIDs, iFlytek Spark's suggestion included safflower oil, which is not practical for oral use.

Moreover, iFlytek Spark differed from both the expert consensus and ChatGPT in areas such as perioperative blood sugar control and preoperative discussions on anticoagulant medication. It frequently offered standard values, resulting in responses that were less clinically relevant. Nevertheless, the model aimed for precision and substance in its responses.

Interestingly, in one of the 35 questions, both AI models deviated from the consensus, which advised administering prophylactic antiemetic medication 30 minutes before the end of surgery. Instead, the models suggested tailoring antiemetic usage based on individual patient assessment, with ChatGPT citing updated guidelines from 2020.[12]

Our thorough examination of all responses indicated that AI responses were generally more organized and comprised detailed analyses, highlighting the significance of patient-specific evaluations and the adherence to hospital and physician protocols. The novelty of the conclusions drawn from this study is presented in the table below (Table 1).

Table 1 Novelty of Study Results

Aspect	Findings	Novelty/Significance
Alignment with Expert Consensus	ChatGPT aligned with expert consensus in 29 out of 35 questions; iFlytek Spark aligned in 22 questions.	Demonstrates AI's potential to closely align with expert consensus in clinical decision-making for ankle fracture management, providing evidence of AI's relevance in healthcare.
AI Response Consistency	Both AI models agreed on 23 questions, but differences were observed in areas like blood pressure management and pain control strategies.	Highlights the ability of AI models to generate consistent recommendations, yet identifies variability in specific medical situations, underlining the complexity of AI's role.
Inconsistencies in Recommendations	ChatGPT suggested blood pressure management within ±20% of normal range, while iFlytek Spark suggested fixed SBP: 120–140 mmHg, DBP: 80–90 mmHg.	Unveils specific inconsistencies between AI recommendations and expert consensus, suggesting a need for further refinement in AI models.
Clinical Relevance	ChatGPT provided more detailed and clinically relevant responses, whereas iFlytek Spark's answers were more standardized and sometimes less applicable.	Indicates that ChatGPT's advanced processing may result in more practical, nuanced clinical advice, while iFlytek Spark's responses may be too generalized.
AI's Role in Personalization	AI demonstrated its ability to provide personalized medical guidance by citing sources and supporting clinical decisions with evidence.	Showcases the potential for AI to offer personalized treatment options, improving decision-making in clinical practice.
Comparative Analysis	Comparison between AI models revealed differences in their alignment with expert consensus, suggesting a gap in AI model development and data.	Highlights the ongoing evolution and refinement needed for AI tools to match clinical expertise fully, especially in areas requiring patient-specific nuances.
Future Applications	The study emphasizes expanding sample sizes and exploring AI's role in complex surgical scenarios and long-term rehabilitation.	Proposes important future research directions, focusing on expanding AI's applicability to complex orthopedic surgeries and rehabilitation strategies.

Based on this comparative examination of 35 questions, we concluded that AI models, particularly ChatGPT, demonstrated superior timeliness, precision, and comprehensiveness in specific decision-making aspects. However, notable variations in development and performance were apparent among the models. ChatGPT notably excelled in this evaluation.

Discussion

Postoperative complications following ankle fractures are generally challenging to treat, especially traumatic arthritis caused by various factors. Many patients may even require ankle fusion surgery as the final treatment option. Although some scholars have achieved good results using minimally invasive fusion techniques to treat traumatic arthritis of the ankle, preventing complications after fractures remains a key concern for clinicians.[13] However, applying ERAS principles in perioperative management provides an important integrated solution. Through this study, we have observed the potential for more standardized treatment approaches in the future for ankle fractures.

The ability of ChatGPT to self-diagnose common orthopedic conditions and provide medical consultation recommendations was analyzed in a recent study. It was found that ChatGPT's accuracy and consistency in diagnosing five prevalent orthopedic conditions varied, suggesting its potential as an initial resource in medical consultation, although with limitations in reliable self-diagnosis.[14] This contrasts with our findings using AI. During the preparatory phase of our study, we observed that when given specific question guidance and structured dialogue, ChatGPT's advice and insights on orthopedic conditions aligned more closely with our clinical experiences. This could be attributed to our use of the GPT-4.0 version, as opposed to the 3.5 version commonly cited in existing literature.

Throughout our research, we endeavored to assess ChatGPT's capacity to summarize and interpret a spreadsheet containing responses to 35 questions. However, its performance in discerning nuances and consistencies fell short of expectations. The model sometimes misconstrued text and did not fully grasp the entirety of the dataset. This limitation persisted even after the file was translated into English, indicating that while ChatGPT excels in language processing, it encounters difficulties in effectively summarizing and synthesizing extensive text data and documents.

In recent years, AI, particularly natural language processing tools like ChatGPT, has demonstrated significant potential in orthopedic practice. According to Giorgino et al (2023), ChatGPT facilitates complex clinical decision-making by enabling the exchange of patient information and providing precise, accessible assistance to both healthcare professionals and patients. Additionally, ChatGPT goes beyond disseminating information; it aids in differential diagnosis, recommends suitable imaging tests, and refines treatment plans based on evidence-based medical guidelines.[15]

The integration of AI into orthopedics, a practice which seems to be particularly prevalent in Asia, improves diagnostic and treatment accuracy, surgical navigation, predictive analytics, and postoperative rehabilitation monitoring (Wang, 2023). These applications not only enhance the precision and efficiency of medical processes but also enable healthcare professionals to interpret complex medical images more accurately, optimizing treatment plans and enhancing patient outcomes. In China, AI technology has led to the development of advanced medical imaging analysis systems. These systems have demonstrated high accuracy in clinical trials and have been instrumental in assisting orthopedic doctors with disease diagnosis and treatment planning (Wang, 2023).[16]

Studies by Kurmis and Ianunzio reveal that deep learning and machine learning significantly enhance surgical planning and execution, leading to improved patient outcomes through optimized data processing and decision-making. AI technology, by analyzing extensive surgical data, helps healthcare professionals make more accurate diagnoses and treatment plans. Its applications in orthopedic surgery include preoperative planning, surgical navigation, and postoperative recovery monitoring. Utilizing deep learning algorithms, AI leverages historical surgical data, to inform decisions on optimal surgical approaches for future procedures. Moreover, AI can analyze data during surgery in real-time, providing immediate feedback that enhances both surgical precision and safety.[17]

Bagaria and Tiwari (2022) emphasize that AI-enhanced robots, equipped with advanced computer vision and sensor technology, perform not only basic operations but also offer real-time feedback during surgeries, thus enhancing surgical accuracy and efficiency. This technology, when combined with healthcare professionals' expertise, significantly optimizes surgical processes and outcomes, particularly in joint replacement surgeries.[18]

Despite the considerable potential of ChatGPT in medical applications, it faces several operational challenges. Ferdush et al noted that during the training phase, ChatGPT may develop biases that could undermine the objectivity and fairness of decision support. Additionally, a lack of profound understanding of complex medical environments may cause ChatGPT's responses to miss necessary contextual relevance, limiting its clinical application. Therefore, continuous human oversight and rigorous evaluation are crucial to ensuring that ChatGPT positively impacts clinical decision-making. While AI tools like ChatGPT can enhance medical decision-making processes, it is crucial not to overlook their limitations and potential ethical issues in practical applications. Future research must address these challenges to ensure the responsible use of AI technology, ultimately realizing its potential to improve the quality and efficiency of medical services (Ferdush et al, 2023).[19] This will not only support technological advancement but also ensure the appropriate application of AI in medical practice.

While AI shows great potential in orthopedics, it's important to recognize its limitations. The accuracy of AI algorithms depends on the quality and quantity of the input data. If the data used is flawed or biased, it can lead to incorrect conclusions. Therefore, AI should be viewed as a complementary tool, requiring validation and oversight by healthcare professionals to ensure its reliability and accuracy.

The integration of AI models like ChatGPT into clinical decision-making presents ethical challenges that must be addressed. This study highlights four key concerns: accuracy and reliability, bias and fairness, transparency and accountability, and data privacy. 1. Accuracy and Reliability: AI-generated medical information is not always accurate, posing risks if used without verification. Chow et al note that AI chatbots can produce misleading yet plausible medical responses, underscoring the need for human oversight. Our findings reinforce this concern, as AI recommendations occasionally diverged from expert consensus. 2. Bias and Fairness: AI models are trained on large datasets that may contain biases, affecting their recommendations. Siddique et al stress that AI should be trained with diverse datasets to reduce bias in clinical applications. Our study found that ChatGPT and iFlytek Spark provided differing responses to certain queries, likely reflecting variations in training data and algorithms. 3. Transparency and Accountability: The "black box" nature of AI makes it difficult for clinicians to assess decision-making processes. Siddique et al emphasize the importance of explainability mechanisms, such as source citations, to improve AI accountability. In our study, ChatGPT provided references, while iFlytek Spark's justifications were less consistent, highlighting the need for greater transparency in AI-driven medical tools. 4. Data Privacy and Security: AI applications in healthcare raise concerns about patient data protection. Chow et al highlight that compliance with regulations such as HIPAA and GDPR is critical for maintaining trust. Although our study did not involve direct patient data, future AI applications in orthopedic care must prioritize secure data management. While AI shows promise in ERAS-guided ankle fracture management, its ethical challenges must be addressed. Future research should focus on bias reduction, transparency, and regulatory compliance to ensure AI's responsible integration into clinical practice.[20,21]

This research comprehensively explores AI's involvement in ankle fracture treatment, yet several inquiries remain unanswered. Subsequent studies should explore AI's capabilities in complex surgeries and improving long-term rehabilitation outcomes. Moreover, there is a need to delve into how AI can seamlessly integrate into physicians' clinical decision-making processes, which represents a crucial area for future research. The rapid advancement of AI is increasingly reshaping healthcare, transforming the medical landscape. ChatGPT, in particular, has quickly emerged as a valuable tool in managing complex orthopedic conditions, offering recommendations comparable to those of experienced orthopedic surgeons.

However, the development and integration of AI encounters significant challenges, including concerns regarding data privacy and security, transparency and interpretability of algorithms, incorporation into clinical practice, and acceptance among physicians and patients. Harnessing the full potential of AI requires interdisciplinary collaboration among healthcare professionals, technology developers, and policymakers. As technology continues to merge with medical practice, AI is poised to play a vital role in personalized medicine, surgical planning, direct surgical assistance, and rehabilitation management. Furthermore, AI advancements are expected to enhance patient education and engagement, thereby improving treatment experiences and satisfaction. The future of AI in orthopedic treatment is promising, potentially ushering in a new era characterized by precision, efficiency, and individualization in the field.

Table 2 Summary of Study Participants

Participant	Role	Years of Experience	Responsibilities
Researcher 1	Trauma Orthopedic Specialist	18 years	Designed the study, completed and evaluated AI responses, participated in data analysis.
Researcher 2	Trauma Orthopedic Specialist	20 years	Evaluated AI responses, participated in data analysis.
Researcher 3	Research Coordinator & Ethics Consultant	10 years	Coordinated study progress, collected data, analyzed data, tracked participants, conducted ethical review.
Researcher 4	Trauma & Microsurgical Reconstruction Specialist	17 years	Collected relevant research data, performed data analysis and comparison.
Researcher 5	Computer Science & Sports Training Expert	10 years	Guided the use of large language models, collected data.

Conclusion

In this study, a comparative assessment of AI models was conducted, specifically ChatGPT and iFlytek Spark, in accordance with expert consensus on ankle fracture treatment protocols guided by ERAS principles. It evaluated the suitability and accuracy of AI in diagnosing and treating ankle fractures. The results indicate that ChatGPT consistently agrees with the expert consensus in most aspects, particularly in providing detailed medical advice and information. However, there were instances of discrepancy, such as managing intraoperative blood pressure and preoperative antiemetic medication, where ChatGPT's recommendations slightly differed from expert opinions, albeit with improved timeliness and specificity. Overall, it offered valuable advice consistent with expert consensus.

A significant aspect of this study was that Chinese is utilized for AI interactions, suggesting that language dependency of large language models like ChatGPT may be irrelevant. In contrast, iFlytek Spark sometimes provided standardized responses that were not entirely applicable to practical clinical scenarios. While AI shows promise in aiding medical decision-making, disparities in performance and relevance persist, underscoring the importance of adhering to medical guidelines and consulting with physicians.

Both ChatGPT and iFlytek Spark demonstrated varying degrees of alignment with expert consensus in addressing ERAS-guided ankle fracture management issues. ChatGPT showed a higher degree of conformity in certain areas, while iFlytek Spark also provided responses in agreement with expert views in other aspects. These variations may reflect differences in training data and algorithms between the AI models, which could influence their performance in clinical decision-making contexts.

The reasons behind the discrepancies between the two AI models and expert consensus are multifaceted. They include differences in the training data of the AI models, their ability to handle complex medical information, and the inherent complexities of the questions themselves. In real-world scenarios, this observation underscores the vital role of medical professionals in AI-supported decision-making, particularly in interpreting and applying insights generated by AI. Therefore, adherence to established guidelines and advice from physicians remains crucial. Overall, AI technology presents a promising option for medical decision-making in ankle fracture treatment, with the potential to enhance and supplement existing expert guidelines. However, its reliability and effective integration require further refinement and optimization. The contributions and role distribution of the study participants are summarized in the table (Table 2).

Research Limitations

1. Limited Number of Expert Evaluators: The expert consensus evaluation was conducted by only two trauma orthopedic specialists, which may limit the generalizability of our findings. Future studies should consider involving a larger and more diverse group of experts to enhance the robustness of the conclusions.

2. Selection of AI Models: Our analysis focused solely on ChatGPT and iFlytek Spark, excluding other potentially relevant AI models. This narrow focus may limit the comprehensiveness of our findings. Future research should include a broader range of AI models to provide a more comprehensive evaluation.
3. Resource Constraints: Due to limitations in time and funding, our study was unable to cover a wider sample, which may affect the generalizability of the results. Future research should consider expanding the sample size to improve external validity.

Despite these limitations, our study provides preliminary insights into the application of AI in ERAS-guided ankle fracture management. We recommend that future research addresses these limitations to further validate and expand upon our findings.

Future Research Directions

Based on the findings of this study, future research should consider expanding the sample size, integrating diverse data types, and conducting comprehensive assessments to validate the effectiveness of AI models across various clinical scenarios.

Data Sharing Statement

The sequence data supporting the findings of this study are available in the Harvard Dataverse repository, accessible via https://doi.org/10.7910/DVN/LTMG6P.

Acknowledgment

We are particularly grateful to all the people who have given us help on our article.

Funding

No external funding received to conduct this study.

Disclosure

The authors declare no conflict of interest.

References

1. Johnson MJ, Kandasamy S, Raspovic KM, et al. Fractures and dislocations of the foot and ankle in people with diabetes: a literature review. *Ther Adv Endocrinol Metab*. 2023;14:20420188231163794. doi:10.1177/20420188231163794
2. Biz C, Angelini A, Zamperetti M, et al. Medium-long-term radiographic and clinical outcomes after surgical treatment of intra-articular tibial pilon fractures by three different techniques. *Biomed Res Int*. 2018;2018:6054021. doi:10.1155/2018/6054021
3. Vanderkarr MF, Ruppenkamp JW, Vanderkarr M, et al. Incidence, costs, and post-operative complications following ankle fracture – a US claims database analysis. *BMC Musculoskelet Disord*. 2022;23:1129. doi:10.1186/s12891-022-06095-x
4. Macera A, Carulli C, Sirleo L, Innocenti M. Postoperative complications and reoperation rates following open reduction and internal fixation of ankle fracture. *Joints*. 2018;6(2):110–115. doi:10.1055/s-0038-1653949
5. Li T, Sun Z, Chai Y, et al. Expert consensus on optimizing ankle fracture treatment protocols under ERAS principles. *Chin J Bone Joint Surg*. 2019;12(01):3–12.
6. Siddique S, Chow JCL. Machine Learning in Healthcare Communication. *Encyclopedia*. 2021;1(1):220–239. doi:10.3390/encyclopedia1010021
7. Kuo RYL, Harrison C, Curran TA, et al. Artificial intelligence in fracture detection: a systematic review and meta-analysis. *Radiology*. 2022;304(1):50–62. doi:10.1148/radiol.211785
8. Zech JR, Santomartino SM, Yi PH. Artificial intelligence (AI) for fracture diagnosis: an overview of current products and considerations for clinical adoption, from the AJR special series on AI applications. *AJR Am J Roentgenol*. 2022;219(6):869–878. doi:10.2214/AJR.22.27873
9. Myers TG, Ramkumar PN, Ricciardi BF, Urish KL, Kipper J, Ketonis C. Artificial intelligence and orthopaedics: an introduction for clinicians. *J Bone Joint Surg Am*. 2020;102(9):830–840. doi:10.2106/JBJS.19.01128
10. Chow JCL, Sanders L, Li K. Generative pre-trained transformer-empowered healthcare conversations: current trends, challenges, and future directions in large language model-enabled medical chatbots. *BioMed Informatics*. 2024;4:837–852.
11. Saugel B, Sessler DI. Perioperative blood pressure management. *Anesthesiology*. 2021;134:250–261. doi:10.1097/ALN.0000000000003610
12. Gan TJ, Belani KG, Bergese S, et al. Fourth consensus guidelines for the management of postoperative nausea and vomiting. *Anesth Analg*. 2020;131(2):411–448.
13. Biz C, Hoxhaj B, Aldegheri R, Iacobellis C. Minimally invasive surgery for tibiotalocalcaneal arthrodesis using a retrograde intramedullary nail: preliminary results of an innovative modified technique. *J Foot Ankle Surg*. 2016;55(6):1130–1138. doi:10.1053/j.jfas.2016.06.002

14. Kuroiwa T, Sarcon A, Ibara T, et al. The potential of CHATGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. *J Med Internet Res*. 2023;25:e47621.
15. Giorgino R, Alessandri-Bonetti M, Luca A, et al. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. *Front Surg*. 2023;10:1284015. doi:10.3389/fsurg.2023.1284015
16. Wang Y. Digital orthopedics in the new AI era: from Asia aspect. *Arthroplasty*. 2023;5(1):61. doi:10.1186/s42836-023-00220-4
17. Kurmis AP, Ianunzio JR. Artificial intelligence in orthopedic surgery: evolution, current state and future directions. *Arthroplasty*. 2022;4(1):9. doi:10.1186/s42836-022-00112-z
18. Bagaria V, Tiwari A. Augmented intelligence in joint replacement surgery: how can artificial intelligence (AI) bridge the gap between the man and the machine? *Arthroplasty*. 2022;4(1):4. doi:10.1186/s42836-021-00108-1
19. Ferdush J, Begum M, Hossain ST, et al. ChatGPT and clinical decision support: scope, application, and limitations. *Ann Biomed Eng*. 2024;52 (5):1119–1124. doi:10.1007/s10439-023-03329-4
20. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. *Front Artif Intell*. 2023;6:1166014. doi:10.3389/frai.2023.1166014
21. Chow JCL, Li K. Ethical considerations in human-centered AI: advancing oncology chatbots through large language models. *JMIR Bioinform Biotechnol*. 2024;5(e64406):e64406. doi:10.2196/64406