

Responses From ChatGPT-4 Show Limited Correlation With Expert Consensus Statement on Anterior Shoulder Instability



Alexander Artamonov, M.D., Ira Bachar-Avnieli, M.D., Eyal Klang, M.D., Omri Lubovsky, M.D., Ehud Atoun, M.D., Alexander Bermant, M.D., and Philip J. Rosinsky, M.D.

Purpose: To compare the similarity of answers provided by Generative Pretrained Transformer-4 (GPT-4) with those of a consensus statement on diagnosis, nonoperative management, and Bankart repair in anterior shoulder instability (ASI). **Methods:** An expert consensus statement on ASI published by Hurley et al. in 2022 was reviewed and questions laid out to the expert panel were extracted. GPT-4, the subscription version of ChatGPT, was queried using the same set of questions. Answers provided by GPT-4 were compared with those of the expert panel and subjectively rated for similarity by 2 experienced shoulder surgeons. GPT-4 was then used to rate the similarity of its own responses to the consensus statement, classifying them as low, medium, or high. Rates of similarity as classified by the shoulder surgeons and GPT-4 were then compared and interobserver reliability calculated using weighted κ scores. **Results:** The degree of similarity between responses of GPT-4 and the ASI consensus statement, as defined by shoulder surgeons, was high in 25.8%, medium in 45.2%, and low 29% of questions. GPT-4 assessed similarity as high in 48.3%, medium in 41.9%, and low 9.7% of questions. Surgeons and GPT-4 reached consensus on the classification of 18 questions (58.1%) and disagreement on 13 questions (41.9%). **Conclusions:** The responses generated by artificial intelligence exhibit limited correlation with an expert statement on the diagnosis and treatment of ASI. **Clinical Relevance:** As the use of artificial intelligence becomes more prevalent, it is important to understand how closely information resembles content produced by human authors.

Anterior shoulder instability (ASI) represents the most prevalent type of shoulder instability, affecting roughly 1% to 2% of the general population.¹ Young, active, and athletic individuals are particularly prone to experiencing events related to shoulder instability.^{1,2}

Recent years have witnessed shifts in decision-making algorithms, indications for surgical intervention, the choice of surgical methods, and the choice of grafts. Management of ASI continues to be controversial, with a lack of international guidelines on various aspects of diagnosis and treatment. This led to a comprehensive consensus statement on the management of ASI published in 2022 by Hurley et al. along with The Anterior Shoulder Instability International Consensus Group.³

The authors used a Delphi consensus methodology and obtained either unanimous or strong consensus on 84% of the questions presented to an international panel of experts. The study was published as a 3-part series and was described as “a Herculean effort.”⁴ The final statement reached a consensus on a wide range of topics relating to ASI management, including diagnosis, nonoperative management, Bankart repair, Latarjet, remplissage, glenoid bone-grafting, revision surgery, rehabilitation and return to play, and clinical follow-up.

Created by OpenAI, ChatGPT is a large language model (LLM) that uses artificial intelligence (AI) to

From the Orthopedic Department, Barzilai Medical Center, Ashkelon, Israel (A.A., I.B.-A., O.L., E.A., A.B., P.J.R.); Ben-Gurion University, Beer-Sheva, Israel (I.B.-A., O.L., E.A., A.B., P.J.R.); Sagol AI Hub at ARC Innovation, Sheba Medical Center, Ramat Gan, Israel (E.K.); and Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel (E.K.).

This study was performed at the Orthopedic Department, Barzilai Medical Center, Ashkelon, Israel.

Received October 5, 2023; accepted February 26, 2024.

Address correspondence to Philip J. Rosinsky, M.D., Orthopedic Department, Barzilai Medical Center, Ashkelon, Israel. E-mail: rosinskymd@gmail.com

© 2024 THE AUTHORS. Published by Elsevier Inc. on behalf of the Arthroscopy Association of North America. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>). 2666-061X/231407

<https://doi.org/10.1016/j.asmr.2024.100923>

produce natural and contextually relevant responses to specific prompts or inputs. This groundbreaking innovation has significantly impacted various fields, ranging from natural language processing and customer service to content generation.⁵ GPT-4 represents the next generation of LLMs in the GPT line, with improved language understanding capabilities in addition to handling real-time and up-to-date information better than ChatGPT. Various applications of ChatGPT in orthopaedics have been explored, including assessing ChatGPT's likelihood of passing the orthopaedic surgery board examination as well as improving patient informed consent.⁶⁻⁹

Opinion leaders in technology have coined ChatGPT as revolutionary and equated it with major technological developments in the history of humanity.¹⁰ In contrast, Sam Altman, the co-founder of openAI, wrote the following: "ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness."¹¹

The underlying principle of GPT-4 hinges on the concept of transformer-based machine learning, particularly the Generative Pretrained Transformer (GPT) architecture.¹² This architecture employs a variant of the transformer model, which uses self-attention mechanisms to make next-word predictions based on the context provided in a piece of text.¹³ For this study GPT-4 was used rather than the more basic version, ChatGPT (GPT-3.5), as the more advanced version has been shown to outperform GPT-3.5 in accuracy. For example, a study comparing passing rates of the United States Medical Licensing Examination demonstrated significantly greater rate of correct answers for GPT-4.¹⁴ Training GPT involves a 2-step process: pretraining and fine-tuning. These processes involve automated steps as well as input by human reviewers to customize and better align the results with human values.^{15,16} These processes are performed in the process of development of GPT, and an in-depth discussion of these are beyond the scope of this study.

ChatGPT and other LLMs may play a future role in several aspects of patient management, including use by surgeons as a professional resource as well as a resource used by patients themselves for education and even requesting actual medical advice. The purpose of this study was to compare the similarity of answers provided by GPT-4 with those of a consensus statement on diagnosis, nonoperative management, and Bankart repair in ASI. The hypothesis was that responses of GPT-4 would align closely with the collective viewpoint of a group of shoulder specialists on managing ASI.

Methods

This study was performed in accordance with the ethical standards in the 1964 Declaration of Helsinki. This study was carried out in accordance with relevant regulations of

the US Health Insurance Portability and Accountability Act (HIPAA). Details that might disclose the identity of the subjects under study have been omitted.

The study "Anterior Shoulder Instability Part I—Diagnosis, Nonoperative Management, and Bankart Repair—An International Consensus Statement"³ was carefully reviewed by us, and key questions and consensus statements were extracted. Subsequently, these questions were presented to ChatGPT. For the purpose of this study, the subscription version of ChatGPT, GPT 4.0, was used. The queries to GPT-4 were performed during May 2023.

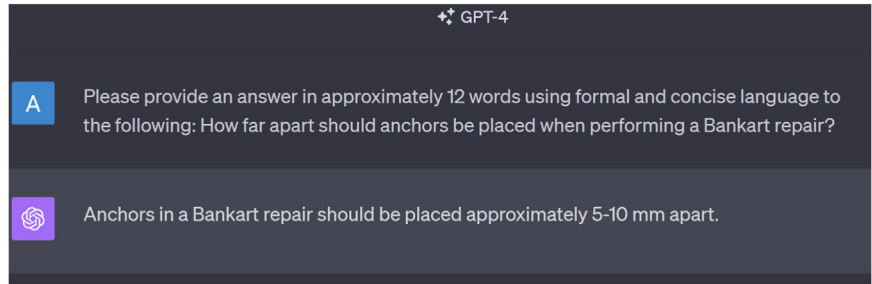
To obtain a clear and easily comparable response, the prompt to GPT-4 was structured as follows: "Please provide an answer in approximately 'X' words using formal and concise language to the following: {text of the question}," where "X" corresponds to the precise word count used in the consensus statement's answer (Fig 1). This adjustment was made due to its tendency to produce excessively lengthy answers, which could potentially hinder the comparison process (Fig 1).

It was noted that when asked multiple times, GPT-4 generated slightly different answers each time. For the purposes of this work, only the first provided answer was used, in order to maintain consistency.

Accurately measuring the similarity between 2 answers using mathematical methods is challenging. Consequently, the use of more subjective measures was considered. Two experienced, fellowship-trained shoulder surgeons were enlisted to assess the answers (P.J.R. and I.B.A.). A table containing questions, consensus statements, and answers generated by GPT-4 was presented to them. The answers were randomized to maintain unawareness among the surgeons regarding which ones were AI-generated and which belonged to the consensus statement. The surgeons were assigned the task of evaluating the similarity of the answers using the terms "low," "medium," and "high" similarity. Surgeons were instructed to designate similarity as "high" if the majority of the answer components were in agreement, "medium" if approximately "half" of the answer components were similar, and "low" when less than half were in agreement. When the surgeons did not agree on the ranking, the opinion of a third shoulder surgeon (E.A.) was requested, and a conclusive decision was made based on the majority opinion.

In addition, the capabilities of GPT-4 to analyze and compare linguistic data were used to compare its own answers with the consensus statement, and then assign a similarity rating—either low, medium, or high to each. GPT-4 was requested to compare its provided response with the answer in the consensus statement using the following query: "Please compare these two answers and evaluate their similarity using only one of the terms: "low," "medium," or "high" similarity:

Fig 1. An example of a query to GPT-4. (GPT-4, Generative Pretrained Transformer-4.)



1. [GPT-4's response]
2. [Consensus Statement]

GPT-4 is known for its ability to use the context provided by previous inquiries and responses to address subsequent questions. In light of this, the conversation history was erased after each request and a new conversation was initiated.

All extracted data were compiled, analyzed, and used to create tables and figures using Microsoft Excel and Microsoft Word (Microsoft Office 2011; Microsoft, Redmond, WA). Weighted κ values were calculated to assess interobserver reliability between the 2 surgeons, and between the agreed surgeon rating and GPT-4 rating.

Results

The summary of questions and answers is presented in [Table 1](#). GPT-4 provided responses for all 31 questions. The degree of similarity, mutually agreed upon by shoulder surgeons, was as follows: high for 8 of 31 questions (25.8%), medium in 14 questions (45.2%), and low for 9 questions (29%). GPT-4, however, assessed the similarity as high in 15 questions (48.3%), medium in 13 questions (41.9%), and low in 3 questions (9.7%) ([Fig 2](#)). The surgeons and GPT-4 reached consensus on the classification of 18 questions (58.1%), whereas a discrepancy occurred on 13 questions (41.9%).

Interobserver reliability for rating of the agreement between the 2 surgeons (P.J.R. and I.B-A.) was 0.33 (fair). Interobserver reliability for rating of the agreement between the surgeons and GPT-4 was 0.41 (moderate).

Discussion

The results of this study demonstrate that GPT-4 responded to questions on the management of shoulder instability with varying consistency when compared with a surgeon consensus statement. The surgeons tended to judge the responses more critically, with 29% of questions being defined as low similarity, whereas GPT-4 only defined 9.7% of questions as low similarity.

Medical students and professionals have started exploring the potential of AI chatbots to assist in

medical studies and to locate answers to contentious subjects within their fields. AI chatbots are promising tools in health care education because of the vast amount of information they can manage.¹⁷ Although AI holds promise for aiding in clinical decision making, the literature offers limited evidence of its effectiveness in this domain.¹⁸ Potential pitfalls associated with GPT-4 use include issues such as the production of incorrect information, potential for bias and discrimination, a lack of transparency and dependability, cybersecurity risks, as well as ethical and societal repercussions.¹⁹

In several of the answers that did not achieve the desired similarity (e.g., questions 7, 19, 21, 27) GPT-4 employed simplistic language, in contrast to the consensus statement, which used intricate terminology essential when delving into the nuances of surgical procedures and diagnostic processes. In these questions the AI demonstrated a lack of comprehension regarding the critical importance of minor details that can significantly impact the diagnosis and treatment of ASI. This could result from differences in potential target end-users; while the consensus statement is geared to expert surgeons, GPT-4 may be more appropriate for the general population.

In specific instances (questions 15, 17, and 18), GPT-4's answers echoed historical approaches to shoulder instability, such as an extended period of immobilization and open Bankart repair, suggesting that the AI might not be fully up to date with recent evidence.^{20,21}

One possible explanation for the dissimilarity of answers provided by GPT-4 could be the training period of GPT-4, which extends until November 2021. However, the bulk of research regarding shoulder instability was published before this cutoff date, implying that the AI should theoretically have had access to relevant information. Of note, the consensus statement used for this study was published after the training period of GPT-4. Other consensus statements that were published before the training period may perform differently, as these could be integrated into the LLM.

GPT-4's assessment diverged from the opinion of the shoulder surgeons in 13 from 31 cases (42%). In the majority of these instances (12 in total), the AI assigned

Table 1. Summary of Questions and Answers, as Well as the Similarity Estimations Provided by Both GPT-4 and Shoulder Surgeons

Question	Consensus Statement	GPT-4 Answer	Shoulder Surgeons Estimation of Similarity	GPT-4 Estimation of Similarity
1. Which aspect(s) of the patient history should be evaluated in the setting of suspected/known primary ASI?	The following should be evaluated: (1) age, (2) gender, (3) mechanism of injury, (4) occupation, (5) sport played and position, (6) level of sport, (7) whether it required reduction, and (8) hyperlaxity.	In cases of suspected/known primary ASI, evaluate the patient's history for previous shoulder dislocations or subluxations, mechanism of injury, frequency of instability episodes, activity level, occupation, and any previous interventions or treatments.	Medium	Medium
2. Which aspect(s) of the patient history should be evaluated in the setting of suspected/known recurrent ASI?	The following aspects of the patient history should be evaluated in the setting of suspected/known recurrent ASI: (1) age, (2) age at first instability event, (3) number of dislocations, (4) original and most recent mechanism of injury, (5) pain, (6) instability symptoms, (7) occupation, (8) sport played and position, (9) level of sport, (10) hand dominance, (11) whether they require reduction/can self-reduce, (12) hyperlaxity, (13) instability with low energy, (14) previous treatment(s), and (15) other injuries/ surgical history.	In the setting of suspected/known recurrent ASI, the following aspects of the patient's history should be evaluated: Age and gender Previous shoulder dislocations or subluxations, including frequency and severity Mechanism of injury for each instability episode Occupation and its potential impact on shoulder health Sports played, position, and level of participation Previous treatments or interventions, such as conservative management or surgeries Hyperlaxity or any underlying connective tissue disorders Functional limitations and impact on daily activities Comorbidities or risk factors that may affect treatment outcomes Assessing these factors will aid in determining the appropriate management strategy for recurrent ASI.	high	high

(continued)

Table 1. Continued

Question	Consensus Statement	GPT-4 Answer	Shoulder Surgeons Estimation of Similarity	GPT-4 Estimation of Similarity
3. Which aspect(s) of the physical examination should be performed in the setting of suspected/known ASI?	The following aspect(s) of the physical examination should be performed in the setting of suspected/known ASI: (1) rotator cuff strength testing, (2) neurologic examination, (3) anterior/posterior apprehension, (4) load and shift, (5) ABER apprehension/relocation, (6) sulcus sign, (7) Beighton score/hyperlaxity of the shoulder at external rotation at side >85, (8) Gagey test, and (9) findings on inspection.	In the setting of suspected/known ASI, perform the following aspects of the physical examination: Inspection for deformities, asymmetry, or muscle atrophy Palpation for tenderness, crepitus, or bony defects Range of motion assessment (active and passive) Strength testing (rotator cuff and deltoid muscles) Specific shoulder instability tests (apprehension, relocation, and anterior release tests) Evaluation of joint laxity and hyperlaxity (sulcus and load-and-shift tests) Examination of the contralateral shoulder for comparison This comprehensive evaluation will help confirm instability and guide treatment decisions.	Medium	Medium
4. In the acute setting, are radiographs required before attempting reduction of a dislocated shoulder? Are postreduction images required?	In the acute setting, radiographs should be performed before attempting reduction of a dislocated shoulder, and postreduction images should be obtained. However, closed reductions can be performed on the field/training room without concern for preradiation radiographs, but postreduction images should still be obtained.	In the acute setting, obtaining preradiation radiographs is recommended before attempting reduction of a dislocated shoulder to identify potential fractures or associated injuries. Postreduction images are also advised to confirm successful reduction and evaluate for any residual abnormalities or complications.	High	High
5. Which plain radiographic views should be obtained to evaluate suspected/known shoulder instability?	The following plain radiographic views should be obtained to evaluate suspected/known shoulder instability (1) anterior-posterior, (2) axillary, and (3) scapular-Y.	To evaluate suspected/known shoulder instability, obtain plain radiographic views including anteroposterior, axillary lateral, and scapular Y-views.	High	High
6. When should advanced imaging (MRI/CT) be performed in a patient presenting with suspected/known ASI?	Advanced imaging should be performed in the following scenarios: (1) irreducible in ED, (2) preoperatively, (3) patient has high risk for recurrence, or (4) suspected rotator cuff injury.	Advanced imaging (MRI/CT) should be performed in patients with suspected/known ASI when evaluating bone defects, assessing soft-tissue injuries, or planning surgical interventions.	Medium	Medium

(continued)

Table 1. Continued

Question	Consensus Statement	GPT-4 Answer	Shoulder Surgeons Estimation of Similarity	GPT-4 Estimation of Similarity
7. Which advanced imaging modality is preferred for a patient presenting with suspected/known ASI, CT, or MRI?	CT should be performed if there is suspected bone loss; otherwise, MRI should be performed.	MRI is the preferred advanced imaging modality for suspected/known ASI, as it better assesses soft-tissue structures.	Low	Medium
8. How should glenoid bone loss be quantified?	Glenoid bone loss should be quantified via the circle method using an enface view of a 3D CT.	Glenoid bone loss can be quantified using 3D computed tomography (CT) scans, comparing the affected glenoid to the contralateral, uninjured side.	Medium	Medium
9. How should humeral bone loss be quantified?	Humeral bone loss should be quantified using 3D CT, and the glenoid track should be evaluated.	Humeral bone loss can be quantified using advanced imaging techniques such as CT scans or 3D MRI for accurate measurements.	Medium	High
10. What limitations should be considered with radiological imaging when extrapolating to expected surgical findings in ASI?	Radiographic approximation of glenoid bone loss may underestimate that identified during surgical evaluation with ASI.	Limitations include potential discrepancies between imaging and actual pathology, and difficulty visualizing soft-tissue injuries or subtle bone defects.	Medium	Medium
11. What are the indications for nonoperative management (of ASI)?	The primary relative indications for nonoperative management include (1) low risk of recurrence, (2) patient's preference to avoid surgery, (3) low functional demand, (4) primary instability, (5) no glenoid bone loss, (6) >30 years old or <14 years old, and (7) timing in-season to allow for return to play.	Indications for nonoperative management of ASI include first-time dislocations, minimal to no bone loss, older patients with lower activity levels, patients unwilling or unable to undergo surgery, and those with medical comorbidities increasing surgical risk. Conservative treatment may involve physical therapy and activity modification.	Medium	High
12. What are the contraindications for nonoperative management (of ASI)?	The primary relative contraindications for nonoperative management include (1) multiple instability events, (2) high risk for further recurrence, (3) severe glenoid bone loss or large bony fragment, (4) instability in low-energy mechanisms, (5) collision athletes, and (6) competitive athletes.	Contraindications for nonoperative management of ASI include recurrent dislocations, presence of bone defects, high-risk sports participation, young age, high-demand occupations, and failure of prior conservative treatments. Surgical intervention may be necessary in these cases.	Medium	High

(continued)

Table 1. Continued

Question	Consensus Statement	GPT-4 Answer	Shoulder Surgeons Estimation of Similarity	GPT-4 Estimation of Similarity
13. What prognostic factors should be considered in patients undergoing nonoperative management (of ASI)?	The prognostic factors that are important to consider specifically in those undergoing nonoperative management include (1) age, (2) athletic demands/ activity level, (3) collision/ contact athlete, (4) number and method of instability events, (5) glenoid bone-loss, (6) Hill-Sachs lesion, and (7) hyperlaxity.	In patients undergoing nonoperative management for ASI, key prognostic factors to consider include (1) age at initial dislocation, (2) frequency of recurrent dislocations, (3) participation in overhead or contact sports, (4) presence of a Bankart lesion, and (5) generalized joint laxity.	Medium	Medium
14. Does immobilization play a role in the nonoperative management of ASI? If so, for what duration of time?	Immobilization may play a role in the early phase of non-operative management of anterior shoulder instability; however, it is unclear how long patients may require to be immobilized.	Immobilization can play a role in nonoperative management of ASI. The duration varies, but typically ranges from 3 to 6 weeks, depending on the severity and individual patient factors.	Medium	High
15. If shoulder immobilization is indicated, should the shoulder be immobilized in neutral, internal, or external rotation?	If patients are immobilized, then they should be immobilized in either neutral or external rotation.	Shoulder immobilization is typically recommended in internal rotation, but optimal position may vary based on individual factors.	Low	Medium
16. When should patients start shoulder range-of-motion exercises (after anterior shoulder dislocation)?	Patients should start shoulder range-of-motion exercises after 1-2 weeks, once comfort permits.	Patients should start shoulder range-of-motion exercises 1-3 weeks after dislocation, depending on severity and physician guidance.	High	High
17. When should patients start resistance training exercises (after anterior shoulder dislocation)?	Patients should start resistance training exercises once full range of motion is restored, and patients can perform the exercises without apprehension.	After anterior shoulder dislocation, patients should initiate resistance training exercises 4-6 weeks after the event, tailored to individual recovery and guided by health care professionals.	Low	Medium
18. When should patients start sport-specific training exercises?	Patients should start sport-specific training exercises once full range of motion and strength are restored, dependent on the timing in the season.	Patients should commence sport-specific training exercises after anterior shoulder dislocation upon receiving clearance from their healthcare professional, typically 4-6 weeks after injury or rehabilitation.	Low	Medium
19. Do corticosteroids or orthobiologics play a role in nonoperative management of ASI?	There is no role for corticosteroids or orthobiologics in nonoperative management.	Corticosteroids and orthobiologics may provide temporary relief in nonoperative management of ASI.	Low	Low

(continued)

Table 1. Continued

Question	Consensus Statement	GPT-4 Answer	Shoulder Surgeons Estimation of Similarity	GPT-4 Estimation of Similarity
20. What are the indications for a Bankart repair?	The primary relative indications for a Bankart repair are (1) primary or recurrent instability, (2) high risk for failure with nonoperative management, (3) minimal glenoid bone loss, (4) on-track Hill-Sachs lesion, (5) patient preference for surgery over nonoperative management, (6) symptomatic instability on examination, and (7) MRI confirmation of labrum tear/Bankart lesion.	A Bankart repair is indicated for patients with recurrent anterior shoulder instability, specifically when a Bankart lesion, a tear of the anteroinferior labrum, is present. Indications include a history of multiple dislocations or subluxations, failed conservative management, high-risk occupations or sports, and a desire to maintain an active lifestyle. Age and degree of joint damage may also influence the decision for surgery.	Low	High
21. What are the contraindications for Bankart repair?	The primary relative contraindications for a Bankart repair are (1) severe glenoid bone loss, (2) off-track Hill-Sachs lesion, (3) uncontrolled epilepsy, (4) posterior instability, (5) multidirectional instability, and (6) likelihood of poor compliance with postoperative rehabilitation.	Contraindications for Bankart repair include significant glenoid bone loss, irreparable rotator cuff tears, multidirectional instability, advanced osteoarthritis, previous failed shoulder stabilization surgery, and medical conditions precluding surgery or anesthesia.	Low	Medium
22. Should Bankart repair be performed arthroscopically or via an open approach? If so, is there an indication for open Bankart repair?	A Bankart repair should be performed arthroscopically. However, an open Bankart repair may be indicated in patients with high risk for recurrence but minimal glenoid bone loss.	Bankart repair can be performed arthroscopically or via open approach, depending on surgeon preference and patient factors. Open Bankart repair is indicated for cases involving extensive tissue damage or complex anatomy.	Low	High
23. Is there an amount of glenoid bone loss above which a Bankart repair should not be performed?	A Bankart repair should not be performed in patients with >15% to 20% glenoid bone loss.	Yes, Bankart repair is not advised when glenoid bone loss exceeds 20% to 25% of the glenoid surface.	Medium	High

(continued)

Table 1. Continued

Question	Consensus Statement	GPT-4 Answer	Shoulder Surgeons Estimation of Similarity	GPT-4 Estimation of Similarity
24. Which prognostic factors should be considered in patients undergoing a Bankart repair?	The following prognostic factors should be considered in patients undergoing a Bankart repair: (1) younger age, (2) glenoid bone loss, (3) Hill-Sachs lesion, (4) male, (5) competitive athlete, (6) overhead-throwing athlete, (7) number of preoperative dislocations, (8) previous shoulder surgery, (9) hyperlaxity, (10) expectations, and (11) ability to comply with postoperative rehabilitation.	Prognostic factors to consider in patients undergoing Bankart repair include age, activity level, occupation, history of dislocations, extent of soft tissue and bone damage, degree of joint hyperlaxity, presence of Hill-Sachs lesions, and the patient's compliance with postoperative rehabilitation. These factors influence the likelihood of recurrent instability and overall surgical success.	High	High
25. What are the indications for performing concomitant procedures with a Bankart repair?	In the setting of a Bankart repair, other pathologies contributing to instability, such as posterior or superior labral tears or Hill-Sachs lesion, should be addressed when performing Bankart repair.	Indications for concomitant procedures with Bankart repair include rotator cuff tears, SLAP lesions, Hill-Sachs lesions, bony defects, and other intra-articular pathologies that require simultaneous treatment.	Medium	Medium
26. Should Bankart repair be performed in the beach/captain's chair or lateral decubitus position?	It is based on surgeon preference whether a Bankart repair is performed in the beach/captain's chair or lateral decubitus position.	Bankart repair can be performed in either beach/captain's chair or lateral decubitus position, depending on surgeon preference and patient factors.	High	High
27. What steps should be taken to minimize complications after a Bankart repair procedure?	Complications, other than recurrence, are rare following a Bankart repair procedure. However, the following can be used to reduce recurrence (1) well-defined rehabilitation protocol, (2) inferior anchor placement, (3) multiple anchor fixation points, (4) small anchors to minimize postage stamp fractures, (5) treatment of concomitant pathologies, (6) careful capsulolabral debridement and reattachment, and (7) appropriate indication and assessment of risk factors.	To minimize complications after a Bankart repair procedure, adhere to the following steps: ensure proper surgical technique, use appropriate patient positioning, achieve sufficient soft-tissue mobilization, avoid overtightening or overcorrection, and closely monitor for infection. Additionally, patient education on postoperative care, rehabilitation, and activity restrictions is crucial. Regular follow-ups and early intervention for complications are also essential to optimize outcomes.	Low	Low

(continued)

Table 1. Continued

Question	Consensus Statement	GPT-4 Answer	Shoulder Surgeons Estimation of Similarity	GPT-4 Estimation of Similarity
28. What is the optimal number of anchors when performing a standard Bankart repair?	A minimum of 3 anchors should be used when performing a standard Bankart repair; however, this may be greater in a more extensive labral tear.	The optimal number of anchors for a standard Bankart repair varies, typically ranging from 3 to 5, depending on the extent of the labral tear and patient anatomy.	High	High
29. How far apart should anchors be placed when performing a Bankart repair?	Anchors should be placed 5-8 mm apart when performing a Bankart repair.	Anchors should be placed approximately 5-10 mm apart during a Bankart repair.	High	High
30. Where should the lowest anchor be placed when performing a Bankart repair in a right shoulder?	The lowest anchor should be placed at 5:30-6:00 o'clock when performing a Bankart repair.	The lowest anchor should be placed at the 5-o'clock position on the glenoid rim in a right shoulder.	Medium	Medium
31. When should a rotator interval closure be performed?	A rotator interval closure may not be routinely recommended but can be considered in those with hyperlaxity.	A rotator interval closure should be performed in cases of multidirectional instability or increased anterior translation without anterior labral damage.	Medium	Low

NOTE. Shown is a comparison of GPT-4.0 answers and consensus statement answers.
 3D, 3-dimensional; ABER, abduction-external rotation; ASI, anterior shoulder instability; CT, computed tomography; ED, emergency department; GPT-4, Generative Pretrained Transformer-4; MRI, magnetic resonance imaging.

a greater similarity rating to its responses. This might suggest that the language model tends to underestimate the intricacies of clinical diagnosis and treatment in relation to patient outcome.

The consensus statement used for this study was designed in a "question-answer" format. Several other consensus statements from various fields of orthopaedics are not designed in a similar fashion, but rather as statements. The ability to use a similar methodology as the one employed for the current study is limited in these cases. However, the current methodology can be replicated for other consensus statements designed as "question-answer" statements.

Aaron Levie, a technology entrepreneur and innovator, is quoted as saying: "ChatGPT will likely play out exactly as innovator's dilemma suggests. To an expert in any given field, it has worse answers. But most people don't have access to experts for everything, so it actually is a productivity boost to everyone else."²² At this point in time, this statement seems accurate regarding the answers on this specific consensus statement questions.

Future iterations of AI language models, and specifically of GPT-based models, may prove beneficial for clinical practitioners. Analyzing a patient's history, examining their clinical findings, and comparing them with published evidence constitutes a task of data analysis, an area in which artificial intelligence could potentially excel. ChatGPT could be trained on a vast

amount of clinical research data and thereby enable clinicians instantly obtain data-driven responses for diagnosis and treatment strategies. This could revolutionize the medical field, making healthcare more efficient, and potentially even more accurate.

It is important to acknowledge that ChatGPT is currently in the early stages of development. Although the promise of these models is exciting, its usage in the medical field, at present, comes with potential dangers. Among these are the issues of "hallucinations" where the AI generates incorrect or misleading information, often with a high degree of confidence. This flaw could lead to serious consequences in a clinical context, where accurate and reliable information is paramount.¹⁷ As such, although the development and application of AI in health care are rapidly advancing, it is critical that caution is exercised, and these models' limitations and potential risks are continually assessed and addressed. As AI capabilities continue to be refined and improved, only the future will determine whether these expectations in the medical field can be fully met. Future development should incorporate safety measures, ethical guidelines, and rigorous validation processes.

Limitations

One primary limitation of this study pertains to the uncertainty regarding the specific information used during the training phase, as well as the variability

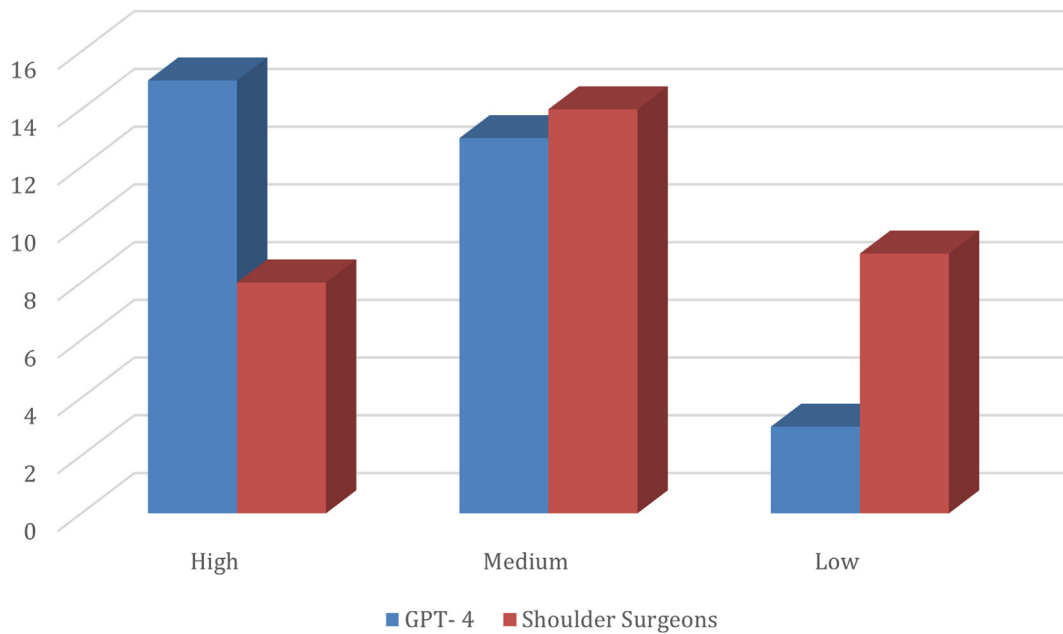


Fig 2. Comparison of estimation of similarity. The diagram shows the estimation of similarity of GPT-4's answers and the consensus statement's answers made by shoulder surgeons and by GPT-4. (GPT-4, Generative Pretrained Transformer-4.)

observed in the responses generated by ChatGPT to each query. In addition, it is important to note that the comparison made in this study between the responses of shoulder surgeons and Chat GPT was qualitative rather than quantitative in nature. This comparison is therefore subjective in nature and agreement was assessed by a Cohen's Kappa score. The agreement in rating of the similarity of the answers was only fair for the involved surgeons and moderate for the consensus rating of the surgeons and ChatGPT. For this reason, the reader is provided with the detailed answers in [Table 1](#), enabling independent evaluation of similarity. Third, according to its developers, GPT-4 surpasses the basic ChatGPT (GPT-3.5) in reliability, creativity, and the ability to manage more intricate instructions. However, they concede that GPT-4 still shares the limitations of its predecessors, most notably, its lack of reliability due to "hallucinations" of facts and reasoning errors. They recommend exercising caution when using language model outputs, especially in high-stakes situations.¹⁶ Lastly, the comparison performed for this study was only regarding a specific part of one consensus statement, and a comparison with other consensus statements may yield different rates of accuracy.

Conclusions

ChatGPT, while capable of aligning with the consensus statement on ASI in a number of statements, exhibits limited correlation in several of the other consensus statements.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work the author(s) used GPT-4 (OpenAI) in order to test the hypothesis of this study. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Disclosures

The authors declare the following financial interests/ personal relationships which may be considered as potential competing interests: P.J.R. is an associate editor for *Arthroscopy* and on the editorial board of *Journal of Arthroplasty*. All other authors (A.A., I.B-A., E.K., O.L., E.A., A.B.) report no conflicts of interest in the authorship and publication of this article. Full ICMJE author disclosure forms are available for this article online, as [supplementary material](#).

References

1. Provencher MT, Midtgaard KS, Owens BD, Tokish JM. Diagnosis and management of traumatic anterior shoulder instability. *J Am Acad Orthop Surg* 2021;29:e51-e61.
2. Sheean AJ, Beer JFD, Giacomo GD, Itoi E, Burkhart SS. Shoulder instability: State of the art. *J ISAKOS* 2016;1: 347-357.
3. Hurley ET, Matache BA, Wong I, et al. Anterior shoulder instability part I—Diagnosis, nonoperative management, and Bankart repair—An international consensus statement. *Arthroscopy* 2022;38:214-223.e7.

4. Lubowitz JH, Brand JC, Rossi MJ. Comprehensive review of shoulder instability includes diagnosis, nonoperative management, Bankart, Latarjet, remplissage, glenoid bone-grafting, revision surgery, rehabilitation and return to play, and clinical follow-up. *Arthroscopy* 2022;38:209-210.
5. Kalla D, Smith N, Samaah S, Kuraku S. Study and analysis of chat GPT and its impact on different fields of study. *IJISRT* 2023;8(3).
6. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. *Clin Orthop Relat Res* 2023;481:1623.
7. Fayed AM, Mansur NSB, de Carvalho KA, Behrens A, D’Hooghe P, de Cesar Netto C. Artificial intelligence and ChatGPT in Orthopaedics and sports medicine. *J Exp Orthop* 2023;10:74.
8. Miika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. *J Bone Joint Surg Am* 2023;105:1519-1526.
9. Kaarre J, Feldt R, Keeling LE, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. *Knee Surg Sports Traumatol Arthrosc* 2023;31:5190-5198.
10. Durantón S. ChatGPT — Let the Generative AI Revolution Begin. Published January 7. *Forbes* 2023. <https://www.forbes.com/sites/sylvainduranton/2023/01/07/chatgpt3let-the-generative-ai-revolution-begin/?sh=327c1c77af15>. Accessed July 20, 2023.
11. Altman S. @sama “ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness.” <https://twitter.com/sama/status/1601731295792414720>. Accessed July 20, 2023.
12. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI blog* 2019;1:9.
13. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 30 Curran Associates, Inc.; 2017;30, 2017. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Accessed July 23, 2023.
14. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. Published online April 12, 2023, <http://arxiv.org/abs/2303.13375>. Accessed September 28, 2023.
15. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33 Curran Associates, Inc.; 2020;1877-1901;33. <https://papers.nips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>. Accessed July 23, 2023.
16. OpenAI. GPT-4 Technical Report. Published online March 27, 2023. <http://arxiv.org/abs/2303.08774>. Accessed July 23, 2023.
17. Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare* 2023;11:887.
18. Kim J-H. Search for medical information and treatment options for musculoskeletal disorders through an artificial intelligence chatbot: Focusing on shoulder impingement syndrome [published online December 19, 2022]. *medRxiv* 2022;12.16.22283512.
19. Borji A. A categorical archive of ChatGPT failures. Published online April 3, 2023. <https://arxiv.org/abs/2302.03494>. Accessed July 20, 2023.
20. Paterson WH, Throckmorton TW, Koester M, Azar FM, Kuhn JE. Position and duration of immobilization after primary anterior shoulder dislocation: A systematic review and meta-analysis of the literature. *J Bone Joint Surg Am* 2010;92:2924-2933.
21. Rashid MS, Arner JW, Millett PJ, Sugaya H, Emery R. The Bankart repair: past, present, and future. *J Shoulder Elbow Surg* 2020;29:e491-e498.
22. Levie A. @levie “ChatGPT will likely play out exactly as innovator’s dilemma suggests. To an expert in any given field, it has worse answers. But most people don’t have access to experts for everything, so it actually is a productivity boost to everyone else.” <https://twitter.com/levie/status/1601683801221971968?lang=en>. Accessed July 20, 2023.