Neurospine 2024;21(1):149-158.
https://doi.org/10.14245/ns.2347052.526

Neurospine
pISSN 2586-6583 eISSN 2586-6591

Original Article

Use of ChatGPT for Determining Clinical and Surgical Treatment of Lumbar Disc Herniation With Radiculopathy: A North American Spine Society Guideline Comparison

Mateo Restrepo Mejia, Juan Sebastian Arroyave, Michael Saturno,
Laura Chelsea Mazudie Ndjonko, Bashar Zaidat, Rami Rajjoub, Wasil Ahmed,
Ivan Zapolsky, Samuel K. Cho

Department of Orthopedic Surgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Corresponding Author
Samuel K. Cho
https://orcid.org/0000-0001-7511-2486

Department of Orthopedic Surgery, Icahn School of Medicine at Mount Sinai, Mount Sinai West, 425 West 59th St, New York, NY 10019, USA
Email: samuel.cho@mountsinai.org

Received: October 13, 2023
Revised: December 5, 2023
Accepted: December 11, 2023

See the commentary on "Use of ChatGPT for Determining Clinical and Surgical Treatment of Lumbar Disc Herniation With Radiculopathy: A North American Spine Society Guideline Comparison" via https://doi.org/10.14245/ns.2448248.124.

Objective: Large language models like chat generative pre-trained transformer (ChatGPT) have found success in various sectors, but their application in the medical field remains limited. This study aimed to assess the feasibility of using ChatGPT to provide accurate medical information to patients, specifically evaluating how well ChatGPT versions 3.5 and 4 aligned with the 2012 North American Spine Society (NASS) guidelines for lumbar disk herniation with radiculopathy.
Methods: ChatGPT's responses to questions based on the NASS guidelines were analyzed for accuracy. Three new categories—overconclusiveness, supplementary information, and incompleteness—were introduced to deepen the analysis. Overconclusiveness referred to recommendations not mentioned in the NASS guidelines, supplementary information denoted additional relevant details, and incompleteness indicated omitted crucial information from the NASS guidelines.
Results: Out of 29 clinical guidelines evaluated, ChatGPT-3.5 demonstrated accuracy in 15 responses (52%), while ChatGPT-4 achieved accuracy in 17 responses (59%). ChatGPT-3.5 was overconclusive in 14 responses (48%), while ChatGPT-4 exhibited overconclusiveness in 13 responses (45%). Additionally, ChatGPT-3.5 provided supplementary information in 24 responses (83%), and ChatGPT-4 provided supplemental information in 27 responses (93%). In terms of incompleteness, ChatGPT-3.5 displayed this in 11 responses (38%), while ChatGPT-4 showed incompleteness in 8 responses (23%).
Conclusion: ChatGPT shows promise for clinical decision-making, but both patients and healthcare providers should exercise caution to ensure safety and quality of care. While these results are encouraging, further research is necessary to validate the use of large language models in clinical settings.

Keywords: Artificial intelligence, ChatGPT, Lumbar disk herniation with radiculopathy, North American Spine Society guidelines, Qualitative study

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © 2024 by the Korean Spinal Neurosurgery Society

INTRODUCTION

Chat generative pre-trained transformer (ChatGPT) is an artificial intelligence (AI) software crafted using a large language model (LLM) and fueled by the GPT-3.5 and -4 model. ChatGPT was trained and developed via supervised learning and reinforcement learning techniques. It uses nondomain specific corpus of textual data from the internet as its main source of information and fine-tunes the accuracy of its responses based on how users rate the response. Since its release in November 2022, ChatGPT has become a popular tool for consulting on a wide array of topics, fine tuning text, and solving questions. The software has demonstrated its high AI reasoning skills with impressive accomplishments like passing the bar exam and the USMLE (United States Medical Licensing Examination) Step Exams.[1] The rise of these AI tools has brought into question their potential use in the field of medicine and more specifically, how they can be incorporated into clinical decision-making in the clinical settings.

Clinical decision-making is heavily aided by clinical guidelines within specific fields and diagnoses. The field of orthopedic surgery has clinical guidelines across the entire pathological spectrum of the field. The Institute of Medicine defines a clinical guideline as follows: "systematically developed statements to assist practitioner and patient decisions about health care for specific clinical situations."[2] Guidelines are especially useful in the medical and surgical management of complex pathologies within orthopedics, specifically those pertaining to the spine. Lumbar disc herniation is a condition that meets these criteria for complex management, and it is also considered the most common cause of lumbosacral radiculopathy.[3,4] The most up-to-date clinical guidelines regarding this condition are the 2012 NASS guidelines.[4] Given that ChatGPT is a relatively new technology deployed to the general public, exploring the usefulness of this technology in predicting the clinical management of various aspects of a pathology is an area of research yet to be explored. Additionally, the user-friendly interface of ChatGPT might encourage patients to use it for medical advice, underscoring the importance of assessing its reliability in offering medically accurate information. Overall, the present study aimed to provide a rigorous evaluation on how well ChatGPT answered questions relating to lumbar disc herniation with radiculopathy based on the NASS 2012 guidelines. Although these guidelines are somewhat outdated and are not the sole resource that physicians use to make their recommendations, they serve as a useful starting point for the management of lumbar disc herniation with radiculopathy. Hence, it is of interest how well ChatGPT's responses matched the 2012 NASS guidelines. The results of this study will have long-standing implications for the use of LLMs in medical decision-making and patient care.

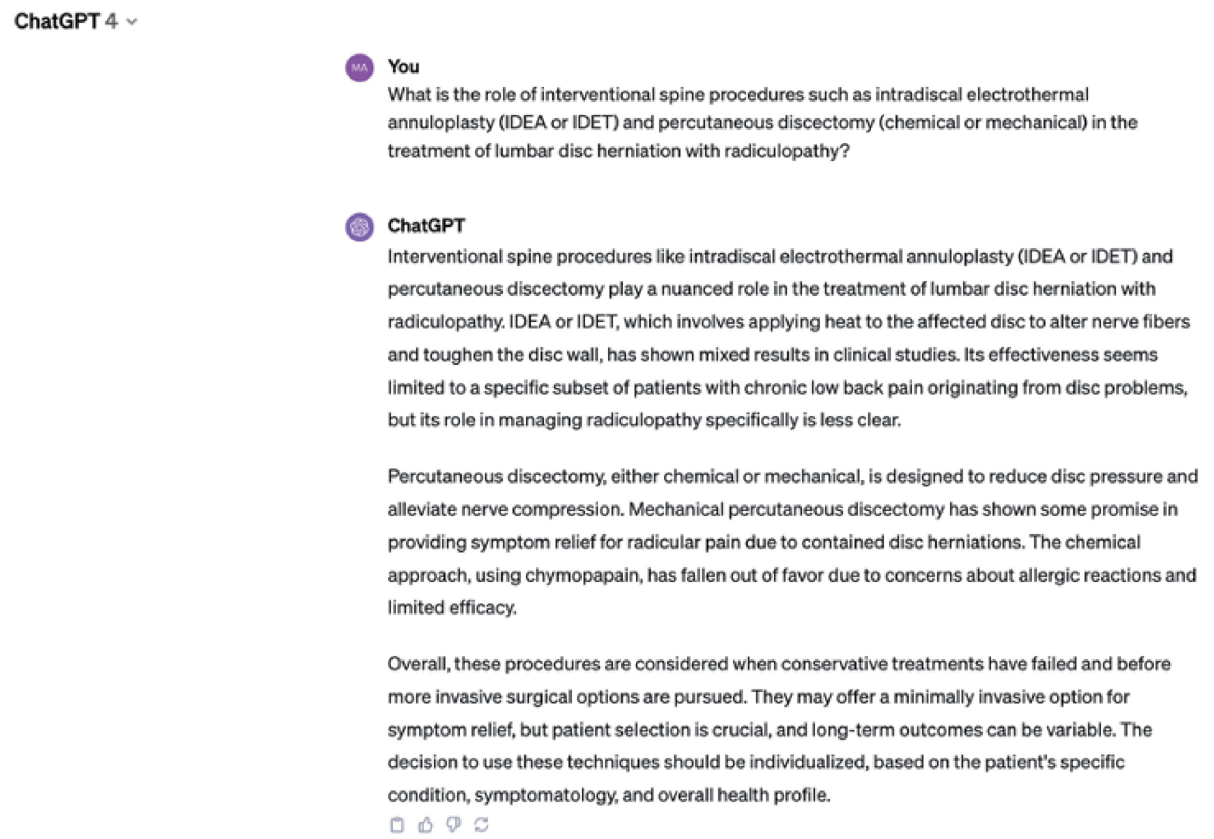
MATERIALS AND METHODS

Institutional review board approval was not needed for this study because ChatGPT is a public resource. The methodology was adapted from previously. To ensure unbiased responses, each question was submitted independently on February 13th, 2023, for the ChatGPT-3.5 version and on October 29th, 2023 for the ChatGPT-4 version without priming, and using a new chat for each question (Figs. 1 and 2).[4] The responses that were obtained from ChatGPT were summarized for readability. In addition, questions that did not explicitly mention spine surgery were modified to include relevant terms to ensure ChatGPT provided specific and targeted responses.

The concordance of ChatGPT responses was evaluated by comparing them to the answers provided by the NASS guidelines under the following 4 criteria: accuracy, overconclusiveness, supplemental, and incomplete. This methodology was adapted from a previously published study from our research group.[5] The grading criteria are described in detail below:

1. Accuracy: Is the ChatGPT response accurate with respect to the NASS guidelines?
   a. If YES, the ChatGPT response did not contradict the NASS guideline.
   b. If NO, the ChatGPT response contradicted the NASS guideline.
2. Overconclusiveness: If the NASS guidelines concluded that there was insufficient evidence to provide a recommendation, did ChatGPT provide one?
   a. If YES, ChatGPT made a recommendation while the NASS guidelines did not provide a recommendation.
   b. If NO, either the NASS guidelines provided a recommendation or both the NASS guidelines and ChatGPT failed to provide a recommendation.
3. Supplementary: Did ChatGPT include additional information relevant to the question which the NASS guidelines did not specify?
   a. If YES, ChatGPT included significant additional information such as references to peer-reviewed articles or further explanations that were not included in the NASS guidelines.
   b. If NO, ChatGPT did not contribute additional information relevant to the question.
4. Incompleteness: If the ChatGPT response was accurate, did ChatGPT omit any relevant details which the NASS guidelines included?
   a. If YES, ChatGPT failed to provide relevant information that was included in the NASS guideline.
   b. If NO, the NASS guideline did not contribute additional information that was not captured by ChatGPT.

https://doi.org/10.14245/ns.2347052.526



You

What is the role of interventional spine procedures such as intradiscal electrothermal annuloplasty (IDEA or IDET) and percutaneous discectomy (chemical or mechanical) in the treatment of lumbar disc herniation with radiculopathy?

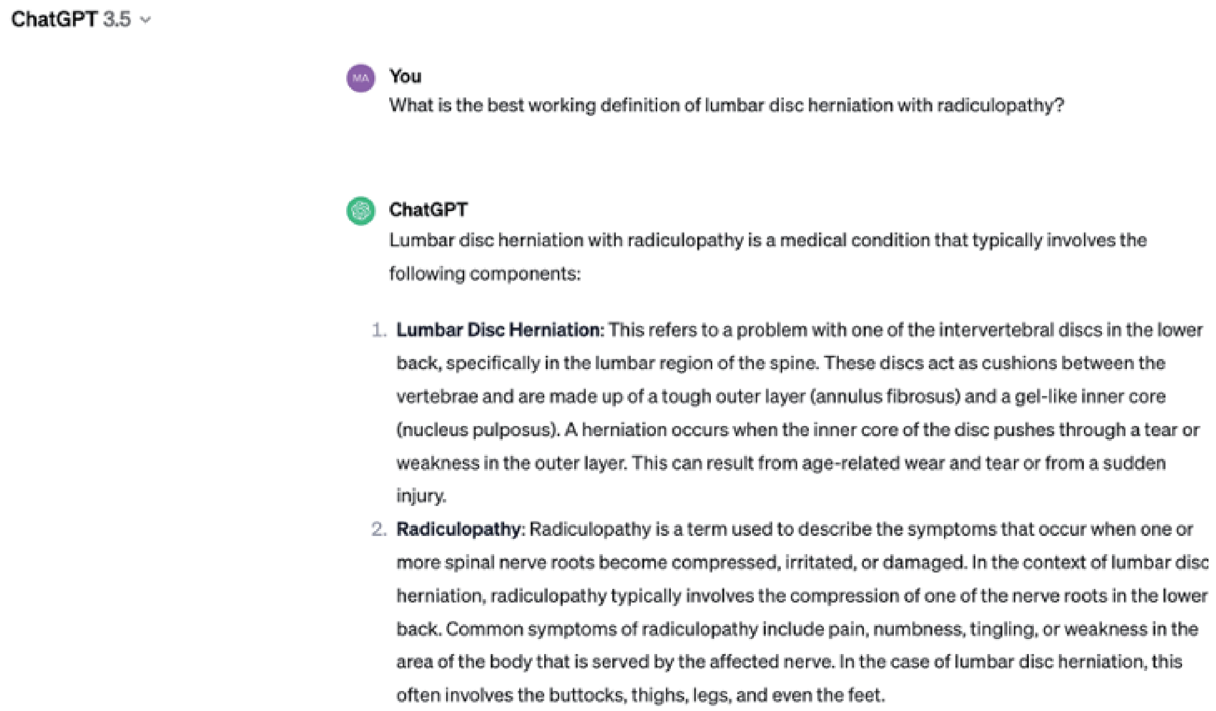
ChatGPT

Interventional spine procedures like intradiscal electrothermal annuloplasty (IDEA or IDET) and percutaneous discectomy play a nuanced role in the treatment of lumbar disc herniation with radiculopathy. IDEA or IDET, which involves applying heat to the affected disc to alter nerve fibers and toughen the disc wall, has shown mixed results in clinical studies. Its effectiveness seems limited to a specific subset of patients with chronic low back pain originating from disc problems, but its role in managing radiculopathy specifically is less clear.

Percutaneous discectomy, either chemical or mechanical, is designed to reduce disc pressure and alleviate nerve compression. Mechanical percutaneous discectomy has shown some promise in providing symptom relief for radicular pain due to contained disc herniations. The chemical approach, using chymopapain, has fallen out of favor due to concerns about allergic reactions and limited efficacy.

Overall, these procedures are considered when conservative treatments have failed and before more invasive surgical options are pursued. They may offer a minimally invasive option for symptom relief, but patient selection is crucial, and long-term outcomes can be variable. The decision to use these techniques should be individualized, based on the patient's specific condition, symptomatology, and overall health profile.

Fig. 1. Example response from ChatGPT-4. ChatGPT, chat generative pre-trained transformer.



You

What is the best working definition of lumbar disc herniation with radiculopathy?

ChatGPT

Lumbar disc herniation with radiculopathy is a medical condition that typically involves the following components:

1. **Lumbar Disc Herniation:** This refers to a problem with one of the intervertebral discs in the lower back, specifically in the lumbar region of the spine. These discs act as cushions between the vertebrae and are made up of a tough outer layer (annulus fibrosus) and a gel-like inner core (nucleus pulposus). A herniation occurs when the inner core of the disc pushes through a tear or weakness in the outer layer. This can result from age-related wear and tear or from a sudden injury.
2. **Radiculopathy:** Radiculopathy is a term used to describe the symptoms that occur when one or more spinal nerve roots become compressed, irritated, or damaged. In the context of lumbar disc herniation, radiculopathy typically involves the compression of one of the nerve roots in the lower back. Common symptoms of radiculopathy include pain, numbness, tingling, or weakness in the area of the body that is served by the affected nerve. In the case of lumbar disc herniation, this often involves the buttocks, thighs, legs, and even the feet.

Fig. 2. Example response from ChatGPT-3.5. ChatGPT, chat generative pre-trained transformer.

The evaluation of ChatGPT's responses was conducted by 2 separate reviewers to confirm the reliability of the grading process. In cases of disagreement, a third author was consulted for resolution. We compared the accuracy, overconclusiveness, supplemental content, and completeness of ChatGPT-3.5 and ChatGPT-4 using a chi-square test, setting the significance level at alpha = 0.05.

RESULTS

There were a total of 29 clinical scenarios that were included in the 2012 NASS guidelines relating to lumbar disc herniation with radiculopathy. ChatGPT-3.5 was accurate in 15 of its responses (52%) in which no contradictions were made to the NASS guidelines. However, ChatGPT-3.5 was overinclusive in 14 of its responses (48%) in which it provided a recommendation when the NASS guidelines did not provide one. In 24 of its responses (83%), ChatGPT-3.5 offered supplemental information, often explaining the procedures involved in various medical interventions or detailing how diagnostic tests are conducted. There were 11 responses (38%) in which ChatGPT-3.5 provided an incomplete response by failing to provide key information that was included in the NASS guidelines. Among the ChatGPT-4 responses 17 (59%) were accurate, 13 (45%) were overconclusive, 27 (93%) were supplemental, and 8 (28%) were incomplete (Table 1, Fig. 3). The differences in accuracy (52%

Table 1. Summary of concordance between NASS clinical guidelines for lumbar disc herniation with radiculopathy and ChatGPT-3.5 and ChatGPT-4 responses

No.	Question
1	What is the best working definition of lumbar disc herniation with radiculopathy?
	ChatGPT-3.5 commentary: The recommendation from ChatGPT-3.5 was accurate, and additionally provided a further explanation describing the mechanism of disc herniation. Accurate: YES, Overconclusive: NO, Supplemental: YES, Incomplete: NO ChatGPT-4 commentary: GPT-4 was inline with the NASS guidelines regarding the definition of lumbar disc herniation with radiculopathy pointing out important factors such as numbness and tingling in a dermatomal distribution. Accurate: YES, Overconclusive: NO, Supplemental: YES, Incomplete: NO
2	What is the natural history of lumbar disc herniation with radiculopathy?
	ChatGPT-3.5 commentary: The model acknowledges that herniated discs can naturally heal with time. Additionally, it incorporates the possibility of requiring surgical intervention and highlights the elevated risk of experiencing additional herniations after the initial occurrence. Accurate: YES, Overconclusive: YES, Supplemental: YES, Incomplete: NO ChatGPT-4 commentary: GPT-4 was accurate in its account that most cases resolve on their own while also highlighting that some cases develop into chronic conditions. Accurate: YES, Overconclusive: NO, Supplemental: YES, Incomplete: NO
3	What history and physical examination findings are consistent with the diagnosis of lumbar disc herniation with radiculopathy?
	ChatGPT-3.5 commentary: GPT provided all of the relevant tests that were recommended in the NASS guidelines such as manual muscle testing, straight leg test, Lasegue's sign, and crosses Lasegue's sign. The model provides further explanations as to how to perform these tests. Accurate: YES, Overconclusive: NO, Supplemental: YES, Incomplete: NO ChatGPT-4 commentary: GPT-4 accurately described the use of a straight leg test as a diagnostic tool but failed to mention Lasegue's sign and crossed Lasegue's sign. It also indicated the loss of deep tendon reflexes as a potential sign of Lumbar Disc Herniation with Radiculopathy which is not backed up by the NASS guidelines. Accurate: NO, Overconclusive: NO, Supplemental: NO, Incomplete: YES
4	What are the most appropriate diagnostic tests (including imaging and electrodiagnostics), and when are these tests indicated in the evaluation and treatment of lumbar disc herniation with radiculopathy?
	ChatGPT-3.5 commentary: GPT accurately provides the recommendation of MRI, CT scan, and electrodiagnostic studies. However, it also included x-ray diagnostics and discography, both of which are not provided in the NASS guidelines. GPT also fails to mention H-reflexes and somatosensory evoked potentials. Accurate: YES, Overconclusive: NO, Supplemental: YES, Incomplete: YES ChatGPT-4 commentary: GPT-4 was accurate in recommending MRI imaging as the gold standard test and the use of CT when MRI is contraindicated. It failed to mention several other tests and their main uses which were described in the NASS guidelines. Accurate: YES, Overconclusive: NO, Supplemental: YES, Incomplete: YES

(*Continued*)

Table 1. Summary of concordance between NASS clinical guidelines for lumbar disc herniation with radiculopathy and ChatGPT-3.5 and ChatGPT-4 responses (Continued)

No.	Question
5	What are the appropriate outcome measures for the treatment of lumbar disc herniation with radiculopathy?
	ChatGPT-3.5 commentary: GPT provided relevant ways to assess positive outcomes for treatment. NASS guidelines provided a link that was accessible. It was therefore not possible to compare answers. Accurate: N/A, Overconclusive: N/A, Supplemental: YES, Incomplete: N/A ChatGPT-4 commentary: GPT-4 provided useful ways to document outcome measures. GPT-4's recommendation could not be compared to NASS due to inaccessibility to the link provided. Accurate: N/A, Overconclusive: N/A, Supplemental: YES, Incomplete: N/A
6	What is the role of pharmacological treatment in the management of lumbar disc herniation with radiculopathy?
	ChatGPT-3.5 commentary: The NASS guidelines explicitly advise against the use of TNF alpha inhibitors, a recommendation which GPT did not acknowledge. While GPT references the majority of treatments cited in the NASS guidelines including glucocorticoids and 5-HT inhibitors, it cautiously suggests that they "may" offer relief, avoiding any absolute assertions. This nuanced approach could be seen as consistent with the NASS guidelines' observations on the insufficient evidence concerning these pharmacotherapies. Accurate: NO, Overconclusive: NO, Supplemental: YES, Incomplete: YES ChatGPT-4 Commentary: GPT-4 makes definitive statements about the use of NSAIDs, oral steroid, muscle relaxants, and neuropathic pain medications while NASS argues that the evidence of these treatments need to be investigated further. Accurate: NO, Overconclusive: YES, Supplemental: NO, Incomplete: YES
7	What is the role of physical therapy/exercise in the treatment of lumbar disc herniation with radiculopathy?
	ChatGPT-3.5 commentary: GPT gives examples of specific low impact exercises that may be used to relieve pain. Given the lack of evidence NASS does not provide specific recommendations but concluded that physical therapy is an option. Accurate: YES, Overconclusive: YES, Supplemental: YES, Incomplete: NO ChatGPT-4 commentary: GPT-4 definitively suggested PT and does not differentiate between the severity of illness. Given the lack of evidence NASS does not provide specific recommendations but concluded that physical therapy is an option. Accurate: YES, Overconclusive: YES, Supplemental: YES, Incomplete: NO
8	What is the role of spinal manipulation in the treatment of lumbar disc herniation with radiculopathy?
	ChatGPT-3.5 commentary: GPT cites the American College of Physicians to make a case on the controversial nature of spinal manipulation; this is not mentioned in the NASS guidelines. Accurate: NO, Overconclusive: NO, Supplemental: NO, Incomplete: NO ChatGPT-4 commentary: GPT-4 posits that the use of spinal manipulation is an option for pain relief which aligns with the NASS recommendations. Accurate: YES, Overconclusive: NO, Supplemental: YES, Incomplete: NO
9	What is the role of traction (manual or mechanical) in the treatment of lumbar disc herniation with radiculopathy?
	ChatGPT-3.5 commentary: GPT highlights the mixed results that have emerged on the role of spinal traction which aligns well with the NASS guidelines lack of recommendation. Accurate: YES, Overconclusive: NO, Supplemental: NO, Incomplete: NO ChatGPT-4 commentary: GPT-4 made definitive suggestions for the use of traction while explaining its potential efficacy. However, NASS posits there is insufficient evidence for or against its use. Accurate: NO, Overconclusive: YES, Supplemental: YES, Incomplete: NO
10	What is the role of contrast-enhanced, fluoroscopic guidance in the routine performance of epidural steroid injections for the treatment of lumbar disc herniation with radiculopathy?
	ChatGPT-3.5 commentary: GPT highlights the appropriateness of this technique and provides useful reasons regarding why it is recommended which aligns with the NASS guidelines. Accurate: YES, Overconclusive: NO, Supplemental: YES, Incomplete: NO ChatGPT-4 commentary: GPT-4 hallucinated references to support its approval of contrast-enhanced fluoroscopy. However, it provided useful reasoning including the increased accuracy of needle placement and reduction of inadvertent intravascular injection. Overall, its indication aligns well with NASS. Accurate: YES, Overconclusive: NO, Supplemental: YES, Incomplete: NO

ChatGPT, chat generative pre-trained transformer; NASS, North American Spine Society; MRI, magnetic resonance imaging; CT, computed tomography; N/A, not available; TNF, tumor necrosis factor; 5-HT, 5-hydroxytryptamine; NSAID, nonsteroidal anti-inflammatory drug.
A full list of the questions and their grading is included in Supplementary Table 2.

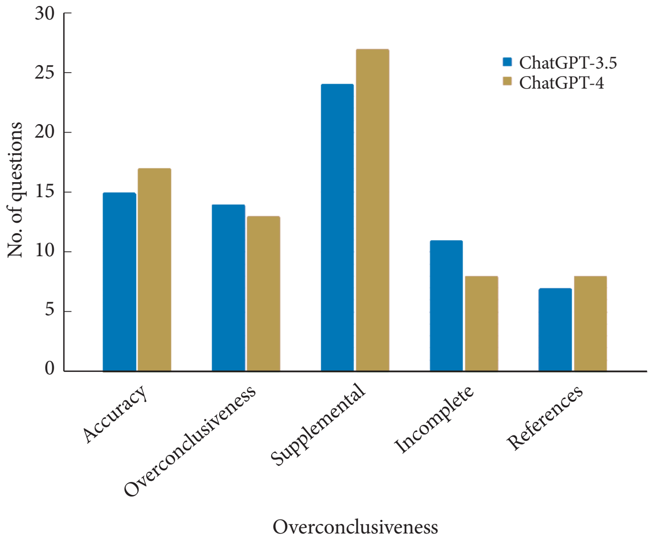


Fig. 3. Accuracy, overconclusiveness, supplementary, and incompleteness of ChatGPT-3.5 and ChatGPT-4 recommendations compared to North American Spine Society clinical guidelines. ChatGPT, chat generative pre-trained transformer.

vs. 59%, p = 0.792), overconclusiveness (48% vs. 45%, p = 1.000), supplemental information (83% vs. 93%, p = 0.650), and completeness (38% vs. 28%, p = 0.313) between ChatGPT-3.5 and ChatGPT-4 were not statistically significant. A detailed overview of the ChatGPT responses, are provided in Supplementary Table 1.

DISCUSSION

Given that the NASS guidelines for lumbar disc herniation with radiculopathy have not been updated since 2012, we anticipate that there may be gaps in the knowledge that were not provided in the text. Specifically, our group investigated some of the questions for which NASS was unable to give recommendations. It was of interest to analyze whether ChatGPT was able to provide more up-to-date information for clinical and patient use.

While no guidelines were published on NASS relating to differences in complication and outcomes between treatment sites, ChatGPT did give recommendations. Among ChatGPT's recommendation, it describes that patients might be at lower risk for infection if surgical treatment for lumbar disc herniation with radiculopathy is treated in ambulatory surgical centers (ASC's) as opposed to hospitals. ASC's have indeed reported lower infection rates compared to hospitals after orthopedic surgery. One study found that the rate of infection in ASC's following orthopedic surgery was 0.33%,[6] while another study from 2021 found that inpatient infection rates following orthopedic surgery at a tertiary hospital was 0.3% for total hip, knee, and shoulder arthroplasty.[7] Interestingly, the same study reported that surgical site infection rates were even lower (0.2%) when the procedures were performed at specialty orthopedic hospitals.[7] This supports ChatGPT's recommendation that there is a difference in complications depending on the surgical site where the procedure is performed.

Another study published in 2013, found that deep infection rates for multispecialty ASCs was 0.81% compared to 0.31% in single specialty ASC's following orthopedic surgery.[8] While the recommendation that ChatGPT gave matched well among orthopedic surgeries in general, our group could not find any sufficient evidence that this is true for lumbar disc herniation with radiculopathy specifically. We hypothesize that ChatGPT generated a recommendation by extrapolating from general infection trends in orthopedic surgery across surgical sites. It is imperative to acknowledge that the incidence of surgical site infections, among other complications and outcomes, can vary considerably depending on the particular type of surgical procedure and the unique circumstances of individual patients. The aforementioned criticality has been duly recognized by ChatGPT.

The NASS guidelines did not provide any information relating to what it considered to be "value of treatment." When ChatGPT was asked this question, it assumed that that the reader was asking about surgical techniques that are used to treat lumbar disc herniation with radiculopathy such as, microdiscectomy, endoscopic discectomy, and open discectomy. Its response stated that microscopic discectomy and endoscopic discectomy are associated with better outcomes and fewer complications compared to open discectomy. A randomized control study published in 2019 found that hospital stay, bone loss, estimated blood loss, and postoperative complications were all lower in microdiscectomy compared to open discectomy for patients with high level lumbar disc prolapse.[9] Although this is in line with ChatGPT's response, the generalizability of this study is limited due to its small sample size.

A meta-analysis from 2022 analyzing pooled randomized control trials found that overall complication rates for full endoscopic lumbar discectomy (FELD) was 5.5% while complications for open discectomy/microdiscectomy was 10.4%.[10] This suggests that patients that received FELD procedures have a lower risk ration for overall complications (risk ratio [RR], 0.55; 95% confidence interval [CI], 0.31–0.98).[10] While this is also in

https://doi.org/10.14245/ns.2347052.526

line with ChatGPT's response, there was no significant difference regarding overall complications between the 2 cohorts. Interestingly, there was a difference in heterogeneous complication results suggesting that patients who receive FELD procedures to treat lumbar disc herniation with radiculopathy are at greater risk for dysesthesia (RR, 3.70; 95% CI, 1.54–8.89), residual fragment (RR, 5.29; 95% CI, 2.67–10.45), and revision surgeries (RR, 1.53; 95% CI, 1.12–2.08).[10] Overall, the literature regarding complications and outcomes between surgical procedures is complex and future studies need to be performed before definitive statements can be made.

The NASS 2012 guidelines for lumbar disc herniation with radiculopathy give no recommendation or have abstained from commenting on the impact of the site-of-service chosen for surgical management on the value of treatment. ChatGPT, however, posits the notion that the chosen site-of-service can affect the value of treatment and defines the latter as "clinical outcomes" and "cost-effectiveness." Despite the lack of evidence that NASS used to justify not commenting on the topic, ChatGPT suggested that hospitals tend to lead to higher costs compared to outpatient settings for the surgical management of this condition. A 2021 study concluded that outpatient lumbar discectomy (a type of surgical management for lumbar herniation) has a higher cost-effectiveness compared to its inpatient counterpart.[11] While this study may support ChatGPT's conclusions regarding cost-effectiveness, this study is not generalizable due to its low sample size (N = 40). Other studies have found that procedures like lumbar microdiscectomy can have reduced costs in the ambulatory setting compared to the hospital associated outpatient centers in patient populations insured with both Medicare and commercial insurances.[12] This evidence may suggest benefits in terms of cost-effectiveness for the surgical management of this spinal condition which is in line with ChatGPT's response. However, these findings may not be generalizable and the patient population for which these findings may be true is a very specific one. The study found that relatively healthier populations would benefit more from outpatient surgery.[13]

It should be emphasized that ChatGPT-3.5 is primarily a language model trained on a broad spectrum of general knowledge rather than specialized medical information,[14,15] which might account for some of the observed variances when compared to NASS guidelines. However, given ChatGPT's ease of access and increasing usage, it was crucial to evaluate its precision in delivering medically accurate information to ensure that patients are receiving reliable data. The latest version of ChatGPT-4, which benefits from training on a more extensive dataset updated as of 2023, has shown potential for improved performance in specialized areas including medical queries. This was evident in a recent comparative study between ChatGPT-3.5 and ChatGPT-4, specifically focusing on their outputs related to thromboembolic prophylaxis in spinal surgery.[5] In our study, however, there were no statistically significant differences between the responses provided by ChatGPT-3.5 and ChatGPT-4 which brings into question ChatGPT-4's ability to perform better at domain specific information. Given this discrepancy, it's crucial to recognize that ChatGPT is an evolving platform and as such, each new version should undergo ongoing validation to analyze its accuracy and reliability so as to ensure that patients are receiving accurate information.

The evolution from ChatGPT-3.5 to GPT-4 represents a significant milestone in the development of advanced language models by OpenAI. ChatGPT-3.5, released in early 2022, was an iteration based on the GPT-3 architecture and boasted around 6 billion parameters. This version was trained on diverse internet text up to 2021, enabling it to generate more contextually relevant and nuanced responses in conversational AI applications. While the exact number of parameters for GPT-4 was not publicly disclosed, it significantly exceeded the 175 billion, and was trained on a dataset inclusive of text and other data types up to the year 2022, enhancing its understanding and generation capabilities across a wider range of contexts and languages.[16] Additionally, OpenAI unveiled ChatGPT-4 Turbo on November 6, 2023, featuring a dataset current as of April 2023. This latest iteration introduces the capability to develop tailored models. With this enhancement, it's possible to input extensive volumes of specialized data for targeted training, leading to more refined and precise outputs.[17] This advancement holds particular significance in medical applications. By feeding the model the latest medical data, specialized in certain areas of healthcare, it's feasible to create a model that provides more accurate and up-to-date medical information.

Before the launch of ChatGPT Turbo-4, various teams had initiated the development of LLMs specifically tailored for the healthcare sector. A prominent example is Med-LLama, introduced in February 2023. This model builds upon the large language model meta-AI (LLaMA) framework, incorporating over 100,000 anonymized conversations between patients and doctors from an online medical consultation service. Med-LLaMA also integrates current information from online resources, along with various offline medical databases. This integration significantly enhances the model's medical expertise and the precision of its advice.[18] Released in 2023, PMC-LLaMA is another significant contribution to the field of medical-specific domain LLMs. As an open-source model, it has undergone extensive refinement, having been fine-tuned with an extensive dataset of 4.8 million biomedical academic papers. This enhancement ensures that both patients and physicians can easily access precise medical knowledge, streamlining the information flow in healthcare settings.[19] It is important to note that literature is lacking on the specific utilization of Med-LLaMA and PMC-LLaMA, highlighting the importance of conduction analyses of these platforms to ensure their safety in the medical setting.

In the medical field, ChatGPT and other LLMs are being explored for various applications. A systematic review conducted following investigated ChatGPT's utility in healthcare education, research, and practice. This review, which included 60 records, identified several benefits of ChatGPT: improved scientific writing, research equity, efficiency in analyzing datasets, utility in drug discovery, streamlining healthcare workflows, cost-saving, enhancing personalized medicine, and improving health literacy. It also found ChatGPT useful in healthcare education for promoting personalized learning and critical thinking. However, 96.7% of the records cited concerns, including ethical and legal issues, risk of bias, plagiarism, inaccurate content, and cybersecurity risks.[20] The review emphasized the need for cautious adoption of ChatGPT in healthcare, highlighting its potential to induce paradigm shifts in the field but also the necessity of addressing its limitations. In another recent study evaluating the diagnostic capabilities of ChatGPT-4, in neuroradiology, 100 cases from the *American Journal of Neuroradiology* were analyzed. ChatGPT's diagnoses, based on patients' medical histories and imaging findings, were compared with published ground truths, achieving an overall accuracy rate of 50%. These findings highlight the potential and limitations of ChatGPT in neuroradiological applications, further emphasizing its varied accuracy depending on specific medical conditions.[21] A recent literature review investigating the use of AI and machine learning in spine surgery, encompassing 46 studies, discovered that AI models achieved an accuracy rate of 74.9%. These models were particularly effective in patient selection, estimating costs, and predicting the length of hospital stays. Additionally, they demonstrated great performance in forecasting functional outcomes and the likelihood of postoperative mortality.[22] This underscores the importance of healthcare personnel to familiarize themselves with AI as a tool to potentially improve patient outcomes.

A significant difficulty in employing ChatGPT and comparable LLMs is their inclination to provide definitive answers to questions, even when a lack of agreement would be more fitting. This phenomenon, called artificial hallucination, happens when these models create plausible yet unsupported statements that are not backed by their training data.[23] ChatGPT-4 was previously found to avoid this issue of fabricated reference production,[5] however, in our study we found that ChatGPT-4 still backed up some of its claims by fabricating studies. In response to queries about references, ChatGPT-4 clarified that it lacked the capability to cite specific references and instead generated responses based on the information it was trained on. Nevertheless, there were instances where ChatGPT-4 did indeed mention valid studies, which were subsequently verified by reviewers. Despite an improved likelihood of providing accurate references, ChatGPT-4 was not immune to occasional instances of generating erroneous information, reflecting the phenomenon of artificial hallucinations.

It is of considerable significance to highlight that the generation of medically pertinent queries is critical in acquiring informative clinical responses from ChatGPT. Although healthcare professionals and individuals with a healthcare-related background possess the necessary expertise to effectively generate such queries, nonexperts may encounter difficulties in doing so when attempting to obtain answers. Nonexperts may pose questions in an inappropriate or imprecise manner, resulting in inadequate responses from ChatGPT if they seek to substitute a medical consultation with a search query. As a result, it is imperative to exercise caution while utilizing ChatGPT and not to regard it as a substitute for medical consultations.

Our study presented some noteworthy limitations. Firstly, the NASS guidelines for lumbar disc herniation with radiculopathy have not been updated since 2012. It is probable that the guidelines have been updated since then with more literature being published on important criteria. This made it difficult to compare NASS guidelines with ChatGPT's generated responses. Additionally, ChatGPT version 3.5 is only able to pull data until 2021, also making it outdated. However, we adjusted for this limitation by comparing ChatGPT's responses to the most recent literature to see if it aligned well with what is being published. Although we thoroughly graded ChatGPT's responses, there exists the potential for subjective bias when determining the grading of each response. To mitigate this, we had 2 independent reviewers' grade ChatGPT's responses while a third reviewer resolved any discrepancies. Despite these constraints, we offer an evaluation of the effectiveness of 2 distinct

ChatGPT models in delivering accurate recommendations for lumbar disc herniation with radiculopathy.

CONCLUSION

ChatGPT shows promising potential for its use in clinical settings. It was often able to provide useful information that was not included in the 2012 NASS guidelines. However, employing ChatGPT as a tool for clinical management should be done so with caution so as to ensure the safety and quality of care given to patients. Although ChatGPT can be used as a tool for answering medically relevant questions, physicians and healthcare workers should continue to refer back to relevant literature, published guidelines, and their own expertise when making clinical decisions. Finally, lay individuals attempting to replace medical consultations with ChatGPT's outputs should also take considerable caution given that its responses are based on the quality of the questions being asked. As AI continues to evolve, further research needs to be performed to investigate its use in clinical settings.

NOTES

Supplementary Materials: Supplementary Tables 1 and 2 can be found via https://doi.org/10.14245/ns.2347052.526.

Conflict of Interest: The authors have nothing to disclose.

Funding/Support: This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution: Conceptualization: MRM, MS, SKC; Data curation: MRM; Formal analysis: MRM, BZ, RR, IZ; Methodology: MRM, JSA, MS, BZ, SKC; Project administration: MRM, BZ, SKC; Visualization: MRM, WA; Writing - original draft: MRM, JSA, MS, LCMN, RR, WA, IZ; Writing - review & editing: JSA, LCMN, BZ, RR, WA, IZ, SKC.

ORCID

Mateo Restrepo Mejia: 0009-0003-0457-3308
Juan Sebastian Arroyave: 0009-0003-9480-0657
Laura Chelsea Mazudie Ndjonko: 0000-0002-3153-4967
Bashar Zaidat: 0000-0002-8823-720X
Rami Rajjoub: 0009-0005-2990-7874
Wasil Ahmed: 0009-0001-0904-1891
Samuel K. Cho: 0000-0001-7511-2486

REFERENCES

1. 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.
2. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the quality chasm: a new health system for the 21st century. Washington (DC): National Academies Press (US); 2001.
3. Tarulli AW, Raynor EM. Lumbosacral radiculopathy. Neurol Clin 2007;25:387-405.
4. Kreiner DS, Hwang SW, Easa JE, et al. An evidence-based clinical guideline for the diagnosis and treatment of lumbar disc herniation with radiculopathy. Spine J 2014;14:180-91.
5. Duey AH, Nietsch KS, Zaidat B, et al. Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations. Spine J 2023;23:1684-91.
6. Edmonston DL, Foulkes GD. Infection rate and risk factor analysis in an orthopaedic ambulatory surgical center. J Surg Orthop Adv 2010;19:174-6.
7. Yayac M, Goswami K, Liss FE, et al. Orthopedic specialty hospitals are associated with lower rates of deep surgical site infection compared with tertiary medical centers. Orthopedics 2021;44:e521-6.
8. Mitchell P, Gottschalk M, Butts G, et al. Surgical site infection: A comparison of multispecialty and single specialty outpatient facilities. J Orthop 2013;10:111-4.
9. Elkatatny AAAM, Hamdy TM, Moenes KM. Comparison between results of microdiscectomy and open discectomy in management of high-level lumbar disc prolapse. Open Access Maced J Med Sci 2019;7:2851-7.
10. Yang CC, Chen CM, Lin MH, et al. Complications of full-endoscopic lumbar discectomy versus open lumbar microdiscectomy: a systematic review and meta-analysis. World Neurosurg 2022;168:333-48.
11. Linhares D, Fonseca JA, Ribeiro da Silva M, et al. Cost effectiveness of outpatient lumbar discectomy. Cost Eff Resour Alloc 2021;19:19.
12. Malik AT, Xie J, Retchin SM, et al. Primary single-level lumbar microdisectomy/decompression at a free-standing ambulatory surgical center vs a hospital-owned outpatient department-an analysis of 90-day outcomes and costs. Spine J 2020;20:882-7.
13. Mikhail CM, Echt M, Selverian SR, et al. Recoup from home? Comparison of relative cost savings for ACDF, lumbar dis-

cectomy, and short segment fusion performed in the inpatient versus outpatient setting. Global Spine J 2021;11(1_suppl):56S-65S.

14. Peng B, Galley M, He P, et al. Check your facts and try again: improving large language models with external knowledge and automated feedback. arXiv:2302.12813v3 [Preprint]. 2023 [cited 2023 Sep 10]. Available from: https://doi.org/10.48550/arXiv.2302.12813.
15. Au Yeung J, Kraljevic Z, Luintel A, et al. AI chatbots not yet ready for clinical use. Front Digit Health 2023;5:1161098.
16. Quintans-Júnior LJ, Gurgel RQ, Araújo AAS, et al. ChatGPT: the new panacea of the academic world. Rev Soc Bras Med Trop 2023;56:e0060.
17. New models and developer products announced at DevDay [Internet]. OpenAI; 2024 [cited 2023 Aug 14]. Available from: https://openai.com/blog/new-models-and-developer-products-announced-at-devday.
18. Li Y, Li Z, Zhang K, et al. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. Cureus 2023;15:e40895.
19. Wu C, Zhang X, Zhang YW, et al. PMC-LLaMA: further finetuning LLaMA on medical papers. arXiv:2304.14454 [Preprint]. 2023 [cited 2023 Aug 14]. Available from: https://doi.org/10.48550/arXiv.2304.14454.
20. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11:887.
21. Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology 2024;66:73-9.
22. Tragaris T, Benetos IS, Vlamis J, et al. Machine learning applications in spine surgery. Cureus 2023;15:e48078.
23. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023;15:e35179.