9-2024

Lumbar Disc Herniation with Radiculopathy: A Comparison of NASS Guidelines and ChatGPT

Ankur Kayastha

Kirthika Lakshmanan

Michael J. Valentine

Anh Nguyen

Kaushal Dholakia

See next page for additional authors

Authors

Ankur Kayastha, Kirthika Lakshmanan, Michael J. Valentine, Anh Nguyen, Kaushal Dholakia, and Daniel Wang

North American Spine Society Journal (NASSJ) 19 (2024) 100333

Contents lists available at ScienceDirect

North American Spine Society Journal (NASSJ)

journal homepage: www.elsevier.com/locate/xnsj

Clinical Studies

Lumbar disc herniation with radiculopathy: a comparison of NASS guidelines and ChatGPT

Ankur Kayastha, BS[a,*], Kirthika Lakshmanan, BS[a], Michael J. Valentine, BA[a], Anh Nguyen, MS[a], Kaushal Dholakia, BS[a], Daniel Wang, DO[b,c]

[a] Kansas City University, Kansas City, MO, United States
[b] MedStar Health, Baltimore, MD, United States
[c] Georgetown University Medical Center, Washington DC, United States

ARTICLE INFO

Keywords:
Artificial intelligence
ChatGPT
Clinical guidelines
Lumbar disc herniation
Radiculopathy
Sciatica
Spine

ABSTRACT

Background: ChatGPT is an advanced language AI able to generate responses to clinical questions regarding lumbar disc herniation with radiculopathy. Artificial intelligence (AI) tools are increasingly being considered to assist clinicians in decision-making. This study compared ChatGPT-3.5 and ChatGPT-4.0 responses to established NASS clinical guidelines and evaluated concordance.

Methods: ChatGPT-3.5 and ChatGPT-4.0 were prompted with fifteen questions from The 2012 NASS Clinical Guidelines for the diagnosis and treatment of lumbar disc herniation with radiculopathy. Clinical questions organized into categories were directly entered as unmodified queries into ChatGPT. Language output was assessed by two independent authors on September 26, 2023 based on operationally-defined parameters of accuracy, over-conclusiveness, supplementary, and incompleteness. ChatGPT-3.5 and ChatGPT-4.0 performance was compared via chi-square analyses.

Results: Among the fifteen responses produced by ChatGPT-3.5, 7 (47%) were accurate, 7 (47%) were over-conclusive, fifteen (100%) were supplementary, and 6 (40%) were incomplete. For ChatGPT-4.0, ten (67%) were accurate, 5 (33%) were over-conclusive, 10 (67%) were supplementary, and 6 (40%) were incomplete. There was a statistically significant difference in supplementary information (100% vs. 67%; p=.014) between ChatGPT-3.5 and ChatGPT-4.0. Accuracy (47% vs. 67%; p=.269), over-conclusiveness (47% vs. 33%; p=.456), and incompleteness (40% vs. 40%; p=1.000) did not show significant differences between ChatGPT-3.5 and ChatGPT-4.0. ChatGPT-3.5 and ChatGPT-4.0 both yielded 100% accuracy for definition and history and physical examination categories. Diagnostic testing yielded 0% accuracy for ChatGPT-3.5 and 100% accuracy for ChatGPT-4.0. Nonsurgical interventions had 50% accuracy for ChatGPT-3.5 and 63% accuracy for ChatGPT-4.0. Surgical interventions resulted in 0% accuracy for ChatGPT-3.5 and 33% accuracy for ChatGPT-4.0.

Conclusions: ChatGPT-4.0 provided less supplementary information and overall higher accuracy in question categories than ChatGPT-3.5. ChatGPT showed reasonable concordance to NASS guidelines, but clinicians should caution use of ChatGPT in its current state as it fails to safeguard against misinformation.

Introduction

Lumbar disc herniation with radiculopathy is a common and debilitating spinal condition that is complex in its presentation, diagnosis, and treatment. Lumbar disc herniations are most prevalent in the third to fifth decade in life with a male to female ratio of 2:1 [1]. The incidence of low back pain is estimated to be between 13% and 31%, but with the addition of radicular symptoms, the incidence ranges

FDA device/drug status: Not applicable.

Author disclosures: ***AK***: Nothing to disclose. ***KL***: Nothing to disclose. ***MJV***: Nothing to disclose. ***AN***: Nothing to disclose. ***KD***: Nothing to disclose. ***DW***: Nothing to disclose.

Statements: The corresponding author has full access to all data in the study and takes responsibility for the integrity and accuracy of this data. This research received no specific grant from any funding agency in the public, commercial, and not-for-profit sectors.

* Corresponding author: Kansas City University, 1750 Independence Ave, Kansas City, MO 64106, USA. Tel.: 816-654-7000.
E-mail address: Kayastha.Ankur@gmail.com (A. Kayastha).

https://doi.org/10.1016/j.xnsj.2024.100333
Received 16 April 2024; Received in revised form 25 May 2024; Accepted 27 May 2024
Available online 1 June 2024

from 12% to 40% for all etiologies [2]. Clinical management of this condition warrants multidisciplinary care and consideration of various prognostic factors, often necessitating the use of established clinical guidelines.

As defined by the North American Spine Society (NASS), lumbar disc herniation with radiculopathy is characterized by localized displacement of disc material beyond the normal margins of the intervertebral disc space resulting in pain, weakness, or numbness in a myotome or dermatome distribution [3]. Isolated lumbar disc herniation refers to the condition where the soft inner core of an intervertebral disc, known as the nucleus pulposus, herniates through the outer fibrous ring of the disc, known as the annulus fibrosus. This pathophysiology ultimately results in disc degeneration, arthritic changes, and subsequent lower back pain due to associated inflammation [1]. The sequela of radiculopathy occurs when herniated disc material compresses or irritates spinal nerve roots exiting intervertebral foramina. This results in potential sensory and motor deficits, paresthesias, and radiating pain. The presentation of symptoms and their severity depends on several factors including location of the herniated disc, extent of nerve compression, and degenerative changes to adjacent anatomy.

Clinical symptoms include low back pain, leg pain, numbness and tingling, muscle weakness, and in more severe cases, bladder and bowel dysfunction [1]. Physical examination may reveal exacerbation of pain in certain positions or weakness along the distribution of lumbosacral nerve roots. Magnetic resonance imaging (MRI) is the gold standard diagnostic modality in confirming lumbar disc herniation [1]. Treatment plans are complex, but they generally span from conservative pain management and physical therapy in mild cases to discectomy or laminotomy in severe cases.

The diagnosis and treatment of lumbar disc herniation with radiculopathy involves comprehensive decision-making processes optimized for individual patients. Clinical guideline references are available for clinician use along with a plethora of streamlined internet-based resources. In the modern era of medicine, artificial intelligence (AI) tools are increasingly being considered to guide clinical decisions due to rapid response rate and ease of access. Many regulatory bodies disseminate clinical guidelines, but there is a fundamental lack of consolidation of this information. AI may help to mitigate this problem.

Chat generative pre-trained transformer (ChatGPT), developed by OpenAI, is one example of a publicly available advanced language model that uses deep learning techniques and large volumes of textual information to produce human-like responses to language inputs [4]. ChatGPT was trained using vast libraries and databases of information, allowing it to predict an answer to a user's question using complex algorithms. AI models have the capability to transform the healthcare landscape through their profound knowledge bases and functional versatility. In an impressive feat, a recent study showed that ChatGPT performed at or near the passing threshold for all 3 United States Medical Licensing Exams (USMLE) without any specialized prompting [5]. ChatGPT demonstrates reasonable accuracy in its general medical knowledge. When presented with clinical vignettes, it can generate differential diagnoses, suggest diagnostic tests, provide final diagnoses, and recommend patient management strategies based on age, gender, and case acuity [6]. The multitude of uses for ChatGPT remains to be explored, but these achievements speak to its potential applications in the medical field.

ChatGPT may play an assistive role in clinical decision-making regarding lumbar disc herniation with radiculopathy. Analyses comparing ChatGPT against NASS guidelines have been conducted for other spine-related conditions, including thromboembolic prophylaxis in spine surgery [7] and lumbar spinal stenosis [8]. In this analysis, the accuracy of ChatGPT responses was compared to The 2012 NASS Clinical Guidelines for the diagnosis and treatment of lumbar disc herniation with radiculopathy, a set of recommendations derived from extensive medical research and expert consensus.

Materials and methods

Workflow

The main objective of this observational study was to assess accuracy and performance of ChatGPT when compared to NASS guidelines. Institutional review board (IRB) approval was not required for this study, as ChatGPT is a publicly available resource, and no patients were involved. The methodology aimed to compare the responses generated by two versions of an AI model, ChatGPT-3.5 and ChatGPT-4.0, to a published set of guidelines called The 2012 NASS Clinical Guidelines for the diagnosis and treatment of lumbar disc herniation with radiculopathy. The original clinical guidelines by NASS were formatted as clinical questions to which answers were supported by evidence-based research. A systematic process was employed to identify relevant clinical guidelines, categorize them appropriately, and evaluate AI-generated responses to these unmodified clinical questions. Responses from ChatGPT were compared to NASS recommendations and graded on operationally-defined parameters: accuracy, over-conclusiveness, supplementary, and incompleteness. To highlight studies referenced in the creation of NASS guidelines, a brief narrative review was conducted for reader supplementation and discussion of ChatGPT concordance.

AI selection

In conceptualizing this project, AI models considered for analysis included Google Bard, Microsoft Bing, and ChatGPT. ChatGPT was ultimately selected for a number of reasons. First, a majority of Americans reported being aware of ChatGPT as a new publicly available AI tool. Approximately 6-in-10 adults, or roughly 58%, of US adults are familiar with ChatGPT [9]. Second, ChatGPT has demonstrated superior efficacy in interpreting and solving case physiology case vignettes when compared to Google Bard and Microsoft Bing [10,11]. This AI is relevant in the current medical literature and has shown promise in being assistive during clinical workflow [5–8,10,11]. Third, the selection of ChatGPT in this study allows for two versions, ChatGPT-3.5 and ChatGPT-4.0, to be directly compared and scored. Comparing an updated version of AI to its predecessor may provide insights on the pace of technological advancements that contribute to increased accuracy and utility of AI responses over time.

Clinical guideline question characteristics and exclusion criteria

A total of twenty-nine clinical guidelines from NASS were initially considered for analysis. Qualitative content analyses were used to generate content categories for the clinical guideline questions. Two authors independently analyzed and assessed the clinical guideline questions to generate content categories. A total of 5 content categories were generated including definition and history, physical findings, diagnostic testing, nonsurgical interventions, and surgical interventions. These categories were adapted from a previous study [8].

To ensure a meaningful comparison between ChatGPT and NASS, a set of exclusion criteria was applied. The questions that were nonconforming to the discussed categories were excluded. For example, questions involving patient functional outcomes or cost-effectiveness of treatment protocols were excluded. Eleven guidelines nonadherent to categories were removed as a result. ChatGPT does not have access to real-time information [4]. This constraint may negatively affect its responses to highly specific and complex questions. Therefore, guideline questions containing multiple subcomponents, comparisons between entities, or references to highly specific or complicated procedures were also excluded. Namely, questions assessing effectiveness of contrast-enhanced fluoroscopic-guided interventional procedures were considered outside the scope of this study. Three additional clinical questions were removed as a result of these criteria. Fifteen clinical guideline questions were included for final data analysis (Fig. 1).

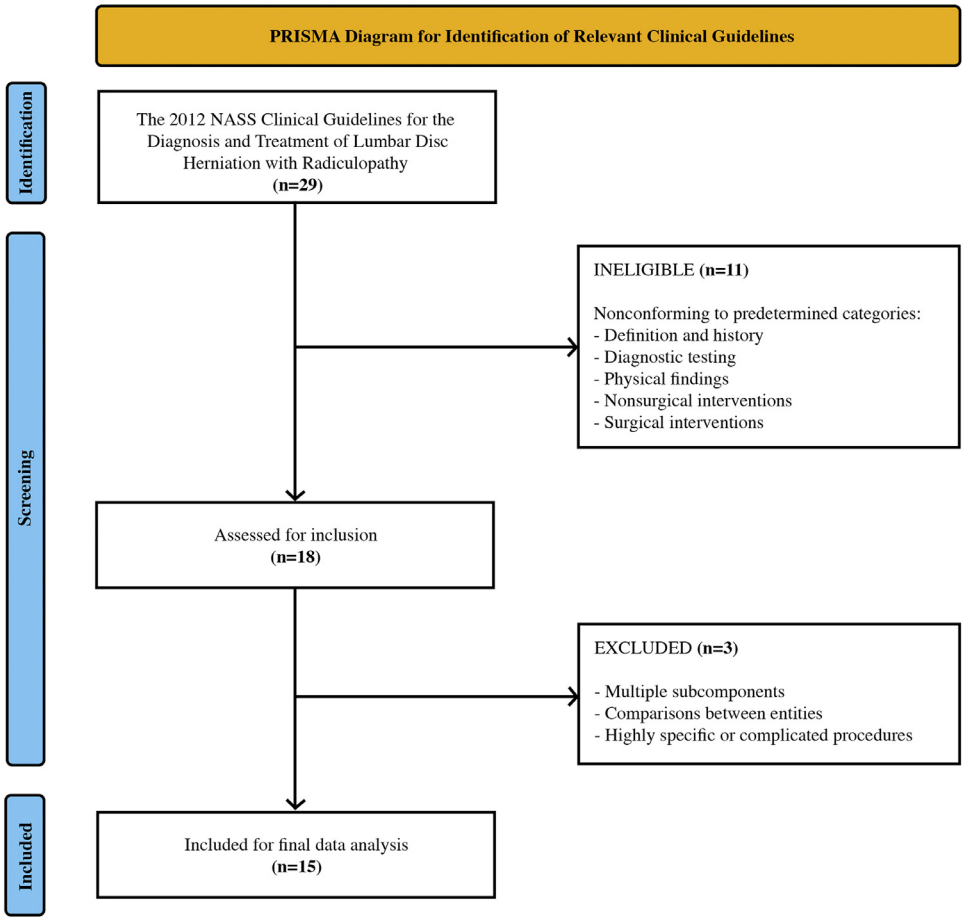

Fig. 1. PRISMA flow diagram illustrating the selection of included NASS guidelines. Exclusion criteria were applied when considering guidelines for analysis.

Outcome measures

ChatGPT responses were evaluated for concordance with NASS guidelines using 4 outcome measures. ChatGPT responses were assessed for accuracy, over-conclusiveness, supplementary, and incompleteness [7]. These measures were adopted from a previous study [7]. Accuracy reflected the overall AI response, while the other measures targeted individual subcomponents of responses. Scoring for outcome measures are further described below:

1. Accuracy: Is the overall ChatGPT response accurate with respect to NASS guidelines?
   a. YES: the overall ChatGPT response did not contradict the NASS guideline.
   b. NO: the overall ChatGPT response contradicted the NASS guideline or provided a definitive recommendation despite insufficient evidence.
2. Over-conclusiveness: If the NASS guidelines concluded that there was insufficient evidence to provide a recommendation on a topic, did ChatGPT provide one?
   a. YES: ChatGPT made a recommendation while the NASS guideline did not provide a recommendation.
   b. NO: either NASS guidelines provided a recommendation or both the NASS guideline and ChatGPT failed to provide a recommendation.
3. Supplementary: Did ChatGPT include additional information relevant to the question which NASS guidelines did not provide or specify?
   a. YES: ChatGPT included a significant amount of additional information or further explanations not included in the NASS guideline.
   b. NO: ChatGPT did not contribute additional information relevant to the question compared to the NASS guideline.
4. Incompleteness: If the ChatGPT response was accurate, did ChatGPT omit any relevant details that NASS guidelines included?
   a. YES: ChatGPT failed to provide relevant information that was included in the NASS guideline.
   b. NO: the NASS guideline did not contribute additional information compared with ChatGPT.

Data collection and analysis

Fifteen clinical guideline questions from The 2012 NASS Clinical Guidelines for the diagnosis and treatment of lumbar disc herniation with radiculopathy were inputted into ChatGPT-3.5 and ChatGPT-4.0 on September 26, 2023. For each guideline question, the prompt was inputted without modifications into a distinct ChatGPT session to prevent biases and ensure no influence from previous inquiries (Fig. 2). If questions did not generate a relevant response, ChatGPT was prompted again with minimal context added to the question (i.e. specifying that the question pertains to lumbar disc herniation with radiculopathy or requesting an answer in paragraph format without bullet points). ChatGPT-generated responses were then recorded verbatim into Microsoft Word alongside NASS recommendations. Transitional phrasing, duplicate information, and unnecessary verbiage in the AI responses were removed. Each ChatGPT-3.5 and ChatGPT-4.0 response was then evaluated by 2 independent reviewers for accuracy, over-conclusiveness, supplementary, and incompleteness compared to NASS guidelines. Discrepancies in scoring were resolved by a third reviewer. Interobserver reliability was calculated using Cohen's kappa coefficient.

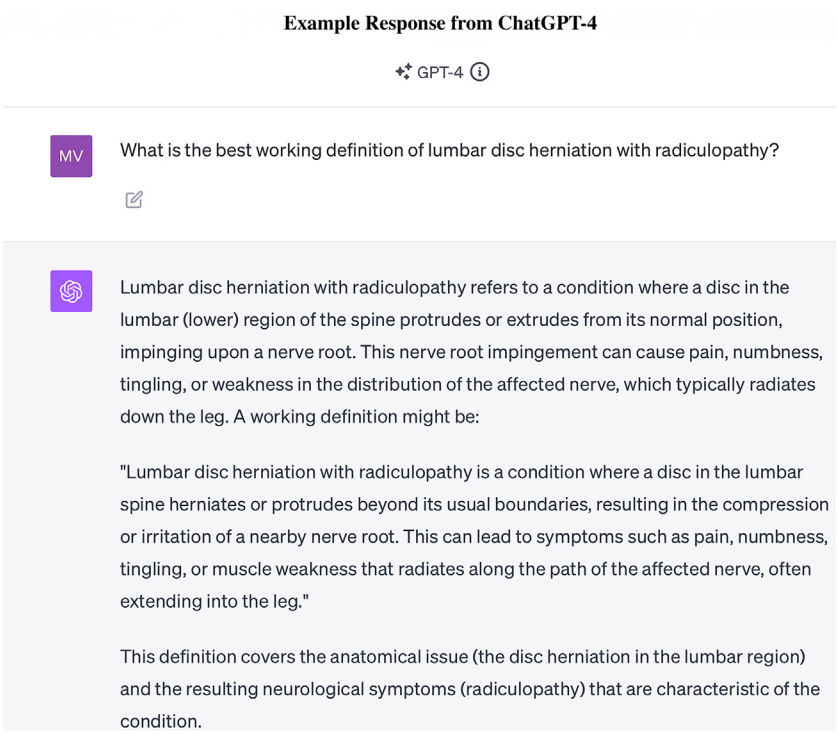

Fig. 2. Example of ChatGPT-4.0 response to a clinical guideline question retrieved from NASS.

This study primarily assessed and compared the language output of ChatGPT to already established information disseminated by NASS. Data were then compiled, analyzed, and visualized in Microsoft Word and Microsoft Excel. ChatGPT-3.5 and ChatGPT-4.0 performance was compared using chi-square tests for the outcome measures. Each chi-square test was conducted using an alpha value of 0.05.

Results

A total of fifteen NASS Clinical Guidelines for the diagnosis and treatment of lumbar disc herniation with radiculopathy were compared against responses generated by both ChatGPT-3.5 and ChatGPT-4.0. Outcome measure scoring for ChatGPT-3.5 and ChatGPT-4.0 responses were detailed (Table 1). A comprehensive list of the NASS guidelines alongside responses from both versions of ChatGPT were recorded (Supplementary Table 1). Among the fifteen responses produced by ChatGPT-3.5, 7 (47%) were accurate, aligning with NASS guidelines without contradictions. Seven (47%) responses were over-conclusive, responding with definitive guidelines where NASS had determined insufficient evidence on the topic. Supplementary information was noted in all fifteen (100%) of its responses, providing additional information beyond NASS guidelines. Six (40%) responses were incomplete, failing to provide relevant details suggested in NASS guidelines. Among the fifteen ChatGPT-4.0 generated responses, 10 (67%) were accurate, 5 (33%) were over-conclusive, 10 (67%) were supplementary, and 6 (40%) were incomplete. The difference in supplementary information (100% vs. 67%; p=.014) between ChatGPT-3.5 and ChatGPT-4.0 was statistically significant. However, accuracy (47% vs. 67%; p=.269), over-conclusiveness (47% vs. 33%; p=.456), and incompleteness (40% vs. 40%; p=1.000) did not show significant differences between ChatGPT-3.5 and ChatGPT-4.0 (Fig. 3). Cohen's kappa coefficient was calculated at a value of 0.966 for interobserver reliability.

The accuracy of responses from both ChatGPT-3.5 and ChatGPT-4.0 across question categories was also analyzed (Fig. 4). ChatGPT-3.5 and ChatGPT-4.0 yielded 100% accuracy for definition and history (two questions) and 100% accuracy for physical examination (1 question) categories. Diagnostic testing (1 question) yielded 0% accuracy for ChatGPT-3.5 and 100% accuracy for ChatGPT-4.0. Non-surgical interventions (8 questions) had 50% accuracy for ChatGPT-3.5 and 63% accuracy for ChatGPT-4.0. Surgical interventions (3 questions) resulted in 0% accuracy for ChatGPT-3.5 and 33% accuracy for ChatGPT-4.0.

Discussion

ChatGPT-3.5 vs ChatGPT-4.0

The results of the comparative analysis showed that ChatGPT-3.5 and ChatGPT-4.0 exhibit differences in performance. ChatGPT-4.0 appeared to be better equipped in accurately responding to questions than ChatGPT-3.5, apparent with higher accuracy percentage in most question categories. These findings align with earlier research on the performance of AI in the medical field. ChatGPT-4.0 has been shown to outperform ChatGPT-3.5 on assessments originally designed for medical students, such as the Medical Knowledge Self-Assessment Program [12]. This difference can be attributed to ChatGPT-4.0 being trained with a substantially larger number of parameters than ChatGPT-3.5, allowing it to comprehend a wider range of queries [13,14]. ChatGPT-4.0 also has larger context windows and a broader base of knowledge due to its updated training data in 2023 as opposed to 2022.

Despite advancements in ChatGPT-4.0, a third of the responses in the data obtained in this comparative analysis still contained inaccurate information. ChatGPT-4.0 outperformed ChatGPT-3.5 in a statistically significant fashion only in terms of one out of four outcome measures: supplementary information. ChatGPT-3.5 added supplemental information to all tested clinical guidelines, whereas ChatGPT-4.0s responses were more conservative. Both models were vulnerable to producing unsupported or irrelevant details, a phenomenon termed "AI hallucination" in the current literature [15]. ChatGPT sometimes generates fabricated data to provide the user with an immediate response, regardless of the content's factual integrity [13,16]. Despite these limitations, ChatGPT may still have potential to be a supplemental source for medical professionals pending future updates and ethical considerations.

Table 1
Summary of outcome measure grading of ChatGPT-3.5 and ChatGPT-4.0.

Category	Question	ChatGPT 3.5				ChatGPT 4.0			
		Accurate	Over-conclusive	Supplementary	Incomplete	Accurate	Over-conclusive	Supplementary	Incomplete
Definition and History	1. What is the best working definition of lumbar disc herniation with radiculopathy?	Y	N	Y	N	Y	N	N	N
	2. What is the natural history of lumbar disc herniation with radiculopathy?	Y	N	Y	N	Y	N	Y	N
Diagnostic Tests	3. 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?	N	N	Y	Y	Y	N	Y	Y
Physical Findings	4. What history and physical examination findings are consistent with the diagnosis of lumbar disc herniation with radiculopathy?	Y	Y	Y	Y	Y	Y	N	Y
Non-Surgical Interventions	5. What is the role of pharmacological treatment in the management of lumbar disc herniation with radiculopathy?	Y	Y	Y	Y	Y	Y	Y	Y
	6. What is the role of physical therapy/exercise in the treatment of lumbar disc herniation with radiculopathy?	N	Y	Y	N	N*	Y	Y	N
	7. What is the role of spinal manipulation in the treatment of lumbar disc herniation with radiculopathy?	Y	N	Y	N	Y	N	N	N
	8. What is the role of traction (manual or mechanical) in the treatment of lumbar disc herniation with radiculopathy?	N	Y	Y	N	Y	N	N	N
	9. What is the role of epidural steroid injections (ESI) for the treatment of lumbar disc herniation with radiculopathy?	Y	N	Y	Y	Y	N	N	Y
	10. Is there an optimal frequency or quantity of injections for the treatment of lumbar disc herniations with radiculopathy?	Y	N	Y	N	Y	N	Y	N
	11. Does the approach (interlaminar, transforaminal, caudal) influence the risks or effectiveness of epidural steroid injections in the treatment of lumbar disc herniations with radiculopathy?	N	Y	Y	N	N	Y	Y	N
	12. What is the role of ancillary treatments such as bracing, electrical stimulation, acupuncture and transcutaneous electrical stimulation (TENS) in the treatment of lumbar disc herniation with radiculopathy?	N	Y	Y	N	N	Y	Y	N
Surgical Interventions	13. Are there signs or symptoms associated with lumbar radiculopathy that predict a favorable surgical outcome?	N	Y	Y	Y	Y	N	Y	Y
	14. When is the optimal timing for surgical intervention?	N	N	Y	Y	N	N	Y	Y
	15. Are there clinical circumstances in which lumbar fusion is appropriate in the treatment of lumbar disc herniation with radiculopathy?	N	N	Y	N	N	N	Y	N

N, no, Y, yes.
* Denotes where 2 separate reviewers held differing assessments of the grading. Disparities were resolved by a third reviewer.

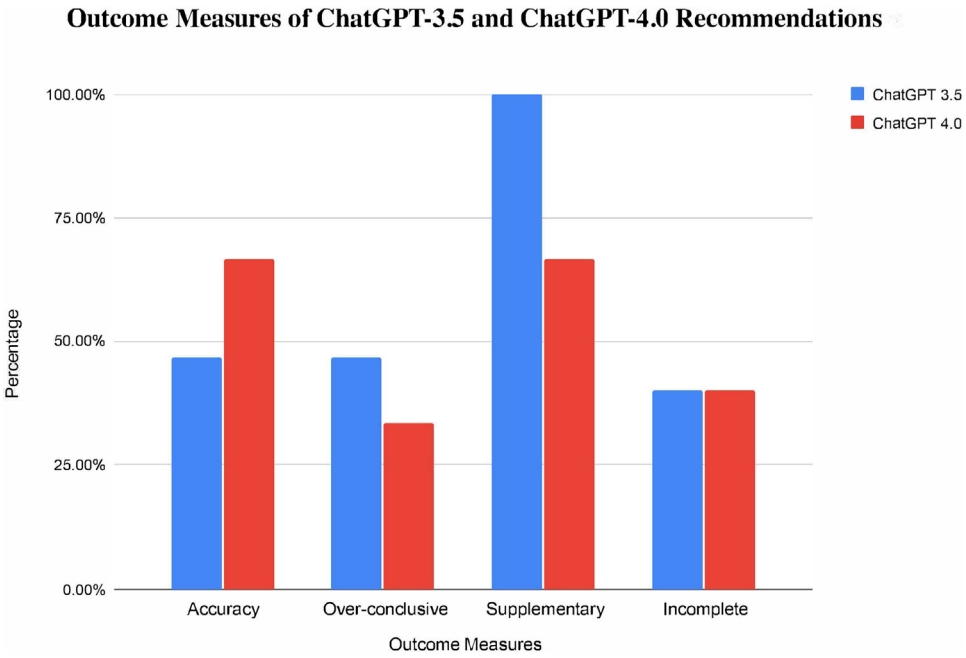

Fig. 3. Accuracy, over-conclusiveness, supplementary, and incompleteness of ChatGPT-3.5 and ChatGPT-4.0 recommendations compared to NASS clinical guidelines.

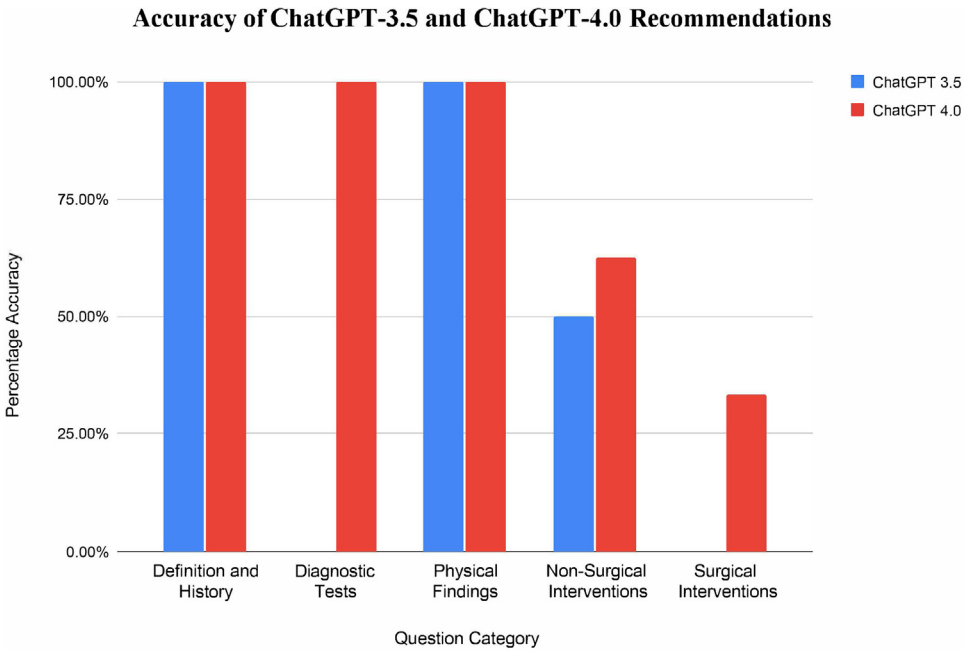

Fig. 4. Accuracy of ChatGPT-3.5 and ChatGPT-4.0 recommendations compared to NASS guidelines stratified by question category.

Assessment of ChatGPT performance

A brief narrative review of The 2012 NASS Clinical Guidelines for the diagnosis and treatment of lumbar disc herniation with radiculopathy was conducted to assess ChatGPT performance in the 5 guideline categories for reader supplementation (Table 2).

Both versions of ChatGPT provided accurate answers for the 2 questions that fell under the category of definition and history for lumbar disc herniation with radiculopathy. The AI models were able to accurately describe the condition, specifying that the intervertebral disc herniates beyond anatomical boundaries and elicits radiating numbness or pain along the path of the affected nerve. Although it did not explicitly specify radiation along myotomal or dermatomal distributions as outlined by NASS, it conveyed that radiation often extends into the leg. Nomenclature and classification of lumbar disc pathology informing these NASS guidelines were based on a literature review conducted in 2001 and revised in 2014.

Herniation was defined as a localized or focal displacement of disc material beyond the limits of intervertebral disc space. Adjacent anatomy includes the lateral recess as the portion of the subarticular zone within the vertebra defined by the medial wall of the pedicle where the same numbered nerve root traverses [17]. L4, L5, and S1 radiculopathies can occur as a result of pressure exerted by the herniated disc onto the thecal sac or lumbar nerve roots. Such pressure may result in nerve root ischemia, inflammation, and characteristic radicular pain down the leg. These symptoms and their correlation with the condition may have shaped ChatGPT's response. Additionally, ChatGPT was correct in describing the natural course of the condition, as 85% to 90% of

Table 2
Brief description of studies referenced in The 2012 NASS Clinical Guidelines for the diagnosis and treatment of lumbar disc herniation with radiculopathy.

Guideline Category	Study (y)	Study Design/Goal	Primary Outcome Assessment	Conclusion	References
Definition and History	Fardon et al (2001, updated 2014)	Literature review conducted to provide a resource that promotes a clear understanding of lumbar disc terminology to clinicians, radiologists, and researchers.	Not applicable	The authors have revised, updated, and provided a widely acceptable nomenclature that helps maintain consistency and accuracy in the description of the anatomic and physiologic properties of the normal and abnormal lumbar disc.	[17]
Diagnostic Tests	Jackson et al (1989)	Prospective cohort study designed to compare the relative accuracies of CT, myelography, CT-myelography, and MRI for the diagnosis of lumbar herniated nucleus pulposus.	Sagittal T1 and T2-weighted images evaluated for neuroradiographic findings from L3-S2	MRI compares favorably to other imaging techniques for diagnosis of lumbar disc herniation due to features including noninvasiveness, lack of ionizing radiation, and high image quality in both sagittal and axial planes.	[18]
	Pape et al (2002)	Prospective cohort study designed to examine the validity of sensory nerve SEP to diagnose sensory radiculopathy in sciatica and to examine whether SEP-diagnosed nerve root compromise is associated with type of radiologically diagnosed degeneration of lumbar spine, presence of sensory sciatic symptoms, spinal level, number of nerve root lesions, previous sciatic episodes, and duration of current episode.	Pathological SEPs as defined by P1 latency or absence of P1 from L4, L5, and S1 nerve roots	The true-positive rate of sensory nerve SEP is higher in patients with facet joint hypertrophy with or without additional disc pathology than in patients with disc pathology only and is higher (although not high enough for screening) when sciatic symptoms are present than absent.	[19]
Physical Findings	Vucetic et al (1996)	Prospective cohort study designed to investigate whether physical signs could predict the degree of hernia (complete hernia, incomplete hernia, protruded disc, and normal disc) found at surgery.	Percentages of Lasegue sign positivity, decreased sensibility, paresis, achilles areflexia, patellar areflexia, scoliosis, and decreased lumbar range of motion.	Stepwise discriminant analysis showed that there were two physical signs of diagnostic value: lumbar range of motion and Lasegue sign. These two signs were able to discriminate between ruptured annuli and intact annuli, impacting choice of invasive therapy.	[20]
	Majlesi et al (2008)	Prospective case control study where both Slump and SLR tests were performed on referred or self-admitted patients with lumbar spine MRI studies demonstrating signs of herniations.	Sensitivity and specificity of SLR and Slump tests.	Both the Slump and SLR tests had similar rates of specificity, but Slump was found to be more sensitive in the study group and subgroups.	[21]
Non-surgical Interventions	Genevay et al (2010)	Randomized clinical trial in which patients with acute severe radicular leg pain and imaging-confirmed lumbar disc herniation received either 2 subcutaneous injections of adalimumab or a matching placebo.	Leg pain defined by VAS. (0-100), lower back pain defined by VAS (0-100), and disability (Oswestry Disability Index)	The addition of a short course of adalimumab to patients experiencing acute severe sciatica resulted in a small decrease in leg pain and back pain, as well as fewer surgical procedures.	[23]
	Bakhtiary et al (2005)	Randomized clinical trial in which patients with clinically diagnosed herniated lumbar disc at L4-L5 or L5-S1 levels were assigned 1 of 2 exercise groups: 4-week LSE protocol then 4-week no exercise (group A), or 4-week no exercise then 4-week LSE (group B).	Pain measurement defined by VAS (0-10), the range of trunk flexion (without pain), the range of left and right SLR, and the time required to complete ADL tasks.	Significant pain relief, left and right SLR angle improvement, increased trunk flexion, and improved ADL performance were observed following exercise periods. LSE protocol may improve ADL performance in patients with HLD.	[22]
	Clarke et al (2007)	Systematic review to determine if traction is more effective than reference treatments, placebo/sham traction, or no treatment for LBP.	Not applicable	Traction cannot be recommended for patients with acute, subacute, and chronic LBP with and without sciatica as a single treatment due to inconsistent results.	[24]
	Ackerman et al (2007)	Randomized clinical trial in which patients were assigned to have LESI therapy every 2 weeks for a total of 3 injections via 3 methods (groups): caudal, interlaminar, or transforaminal.	Numeric pain intensity scores (0-10), Oswestry Low Back Pain Scale (0-70), and Beck depression scores (0-63).	The transforaminal route of epidural steroid placement is more effective than other routes due to more patients reporting complete or partial pain relief. A higher incidence of steroid placement in the ventral epidural space is observed when this method is used.	[25]

(*continued on next page*)

Table 2 (*continued*)

Guideline Category	Study (y)	Study Design/Goal	Primary Outcome Assessment	Conclusion	References
Surgical Interventions	Kohlboeck et al (2004)	Prospective cohort study designed to evaluate the predictive value of medical, general, psychological, pain-related cognition, and coping strategy factors in predicting health outcomes after lumbar discectomy.	Lasegue sign, postoperative pain maintenance and intensity (0-10), pain locations, functional capacity (HMQ), return to work, and health-related quality of life (MOS SF-36).	Lasegue sign, depression, and sensory pain descriptions proved to be significant predictors of outcome, whereas pain cognition and pain coping strategies had no significant influence on outcomes.	[27]
	Peul et al (2007)	Randomized clinical trial to determine whether a strategy of early decompressive HLD surgery leads to better outcomes during the first year than does a strategy of conservative treatment for an additional six months followed by surgery for patients who do not have improvement.	Roland Disability Questionnaire for Sciatica, VAS for leg pain (100 mm), and 7-point Likert self-rating scale of global perceived recovery.	One-year outcomes were similar for patients assigned to early surgery and those assigned to conservative treatment with eventual surgery, but rates of pain relief and perceived recovery were faster for those with early surgery.	[26]

ADL, activities of daily living, CT, computed tomography, HLD, herniated lumbar disc, HMQ, hannover mobility questionnaire, LBP, low back pain, LESI, lumbar epidural steroid injections, LSE, lumbar stabilizing exercises, MOS SF-36, medical outcomes study short form health survey, MRI, magnetic resonance imaging, SEP, somatosensory evoked potential, SLR, straight leg raise, VAS, visual analog scale.

cases are short-lived and regress over time [1]. Both versions of ChatGPT performed best in this category, perhaps due to the lack of clinical judgment required to answer definition-based questions.

One clinical guideline question conformed to the diagnostic testing category. NASS describes that there is relative paucity in high quality studies on advanced imaging. However, the NASS workgroup opinion was that in patients with a history and physical examination findings consistent with lumbar disc herniation with radiculopathy, magnetic resonance imaging (MRI) is considered the most appropriate test for diagnosis [3]. MRI compares favorably to other imaging techniques due to noninvasiveness, lack of ionizing radiation, and high image quality in sagittal and axial planes [18]. Computed tomography (CT) scans may also be appropriate in certain cases, but there is no mention of X-ray utilization within NASS guidelines. ChatGPT-3.5 asserted that X-ray may be used for preliminary imaging, which is incorrect, but ChatGPT-4.0 correctly mentioned its lack of utility in aiding diagnosis.

NASS also posits that while somatosensory evoked potentials (SEPs) can complement cross-sectional imaging, they lack specificity at the level of nerve roots. Pape et al examined validity of sensory nerve SEPs as a diagnostic tool for sensory radiculopathies [19]. The results indicated that the true-positive rate of sensory nerve SEP is higher in patients when sciatic symptoms are present rather than absent [19]. Both versions of ChatGPT failed to mention electrodiagnostics in their responses, leading to a grading of incompleteness.

Physical findings, consisting of one question, was a category that included both history and physical examination findings. ChatGPT-4.0 was much more conservative in its response than ChatGPT-3.5. NASS guidelines emphasized use of Lasegue's sign, straight leg raise, and muscle/sensory testing as recommendations for diagnosis, whereas ChatGPT also mentioned potential diagnostic validity of lumbar range of motion and reflex strength. NASS indicated insufficient evidence for lumbar range of motion and reflex strength. Concurrent with NASS, ChatGPT also reflected a sense of ambiguity. Both versions of ChatGPT did not make definitive claims and were therefore graded as accurate. Vucetic et al employed a stepwise discriminant analysis revealing Lasegue sign and lumbar range of motion as two signs capable of discriminating between ruptured and intact annuli, ultimately impacting choice of invasive therapy [20]. Slump and straight leg raise tests also showed diagnostic validity in patients demonstrating signs of herniations [21]. ChatGPT-3.5 discussed various physical symptoms and imaging modalities irrelevant to the guideline question and NASS recommendations, whereas ChatGPT-4.0 provided a focused response discussing importance of muscle strength and sensory symptoms in concordance with NASS.

The most extensive category was nonsurgical interventions which comprised eight questions. According to NASS, all answers within this category contained some form of insufficient evidence. Both versions of ChatGPT answered approximately half of these questions accurately. NASS recommendations discussed use of TNF-alpha inhibitors, exercise, spinal manipulation, and epidural steroid injections (ESIs) as forms of nonsurgical interventions with varying evidence grades. Most recommendations provided were based on randomized controlled trials consisting of patients with clinical or imaging-diagnosed lumbar disc herniation. Exercise was recommended in the form of workgroup opinion. In one NASS-referenced study, it was found that exercise led to significant pain relief and improved performance in activities of daily living [22]. In terms of medication therapy, a short course of adalimumab (TNF-alpha inhibitor) was shown to decrease leg pain and back pain in patients affected by sciatica [23]. Owing to inconsistent results, spinal traction could not be recommended by NASS [24]. Finally, 1 study showed transforaminal epidural steroid injections were considered superior to both caudal and interlaminar routes. This superiority was attributed to higher levels of complete or partial pain relief in comparison to the other routes [25]. Of note, this injection route was not substantiated as superior by NASS as a whole [3].

Both versions of ChatGPT often made incorrect and definitive claims in this category, recommending use of nonsteroidal inflammatory drugs (NSAIDs), muscle relaxants, and oral corticosteroids as pharmacologic interventions. ChatGPT described physical therapy and exercise as essential components to treatment, perhaps due to the importance of activity as a lifestyle modification in many other spinal conditions. However, ChatGPT was able to appropriately capture the controversial nature of evidence in spinal manipulation and traction in scientific literature. Both versions of ChatGPT also correctly identified the efficacy of epidural steroid injections but utilized unsubstantiated claims to differentiate risks and effectiveness between various injection routes.

Surgical interventions were the last category addressed, containing 3 questions. Surgical outcomes and optimal surgical timing were discussed as topics within related guidelines. ChatGPT performed worst in this category, with ChatGPT-3.5 being incorrect on all questions and ChatGPT-4.0 answering only one correctly. NASS asserts that earlier surgery (within 1 year) was associated with favorable surgical outcomes due to faster rates of pain relief and perceived recovery [26]. Additionally, while Lasegue sign, depression, and sensory pain descriptions

were adequate predictors of surgical outcome, pain cognition, and coping strategies were not [27]. ChatGPT incorrectly asserted that pain, functional impairment, failure of conservative treatments, and MRI-confirmed nerve compression predict favorable surgical outcomes. The AI also stated that there is no universal optimal timing for surgery, directly contradicting NASS. Notably, ChatGPT did provide a caveat and suggested evaluation by healthcare providers for accurate prognosis. Although this was a responsible addition to the overall response, this was not considered in grading due to lack of utility in clinical scenarios.

ChatGPT performance was scrutinized appropriately during grading. Although impressive in its knowledge base and fast rate of response, both versions of ChatGPT often made uncorroborated claims that were not evidence-based. ChatGPT is currently unable to access databases of medical literature, which may partially explain its inaccuracies. In its current state, ChatGPT cannot be recommended as an assistive resource in clinical workflow. If ChatGPT is updated to prevent fabrication and enable controlled access to current recommendations disseminated by regulatory bodies, it may have potential in assisting clinical decision-making.

ChatGPT in spine literature

Prior studies have explored the efficacy of ChatGPT in clinical workflow and its ability to make recommendations regarding other spinal conditions. Duey et al conducted a study on ChatGPT's effectiveness at providing recommendations for thromboembolic prophylaxis in spine surgery [7]. ChatGPT proved to be reasonably accurate when responses were compared to NASS guidelines with ChatGPT-4.0 demonstrating an accuracy of 92%, while ChatGPT-3.5 performed with an accuracy of 33%. The AI models were observed to exhibit several limitations regarding vagueness in response, lack of completeness in clinical planning, erroneous citation of sources, and providing recommendations when NASS guidelines deemed evidence to be inconclusive [7].

Another study conducted by Rajjoub et al explored ChatGPT capabilities in diagnosis and treatment of lumbar stenosis [8]. The authors concluded that the AI provided more specific results when compared to NASS guidelines. For instance, ChatGPT provided specific physical therapy exercises and commented on the effects of ancillary treatments on lumbar stenosis. Additionally, it spoke on aspects of bracing, electrical stimulation, and acupuncture, even expanding on pharmacological treatments while NASS guidelines made no such recommendations [8]. It is important to note that the latter study only conducted data analysis utilizing ChatGPT-3.5, whereas the former study conducted their analysis on both ChatGPT-3.5 and ChatGPT-4.0 to provide comparison between versions of AI.

Continued research on ChatGPT's abilities and limitations is necessary to develop a better understanding of its clinical utility before any consideration is given to its incorporation into clinical workflow. Rao et al addressed this notion by analyzing ChatGPT responses to 36 different clinical vignettes [6]. The authors concluded that ChatGPT improves proportionally in clinical decision-making as more clinical information is provided. In other words, the AI was most effective at clinical planning with respect to final diagnosis and specific clinical presentation as opposed to initial or differential diagnoses [6,28].

Ethical considerations

A full discussion of ethical considerations is beyond the scope of this article; however, it is vital to account for the ethics of integrating AI into healthcare. The widespread nature of AI requires strong privacy protections to keep patient information safe, making the transparency of AI usage in clinical settings paramount [29,30]. Additionally, AI algorithms cannot account for many multifactorial components associated with health, and to prevent disparities in patient care, it is crucial to rectify any inherent biases within AI algorithms [30]. Accountability mechanisms must be established as AI should not serve as a replacement for physician care [29]. If AI is to be effectively employed as a supplementary tool in healthcare settings, it is essential that such models are not only trained extensively in medical knowledge but are also calibrated with stringent ethical and privacy standards. The development of AI systems that adhere to these principles represents a promising direction for future exploration.

Limitations

This study contains limitations that warrant consideration. ChatGPT performance in generating responses was altered by specificity of the questions posed to it and its adaptation to individual user questions at a timepoint. This made it difficult to assess the consistency of AI responses. While the NASS guidelines provided a broad framework, more precise, and specific inquiries tended to yield improved responses [31]. Of note, ChatGPT was graded in all four outcome measures based on a reference document which has not been revised since 2012. ChatGPT-3.5 and ChatGPT-4.0 were trained with information until 2022 or 2023 respectively, and therefore may have led to discrepancies when compared to the NASS guidelines [4,32]. Scoring of outcome measures may have been influenced by a limited number of observers and observer bias. Finally, clinical guideline categories were formulated subjectively and contained a small sample size which may have affected statistical significance and generalizability of the results.

Conclusions

ChatGPT was able to provide recommendations regarding lumbar disc herniation with radiculopathy with reasonable concordance to NASS guidelines. ChatGPT-4.0 demonstrated overall higher accuracy than ChatGPT-3.5 in question categories. ChatGPT-4.0 provided less supplementary information than ChatGPT-3.5, a statistically significant finding. ChatGPT shows potential as an adjunctive resource to clinical workflow, but more validation research and AI updates are necessary before implementation into the healthcare landscape can be recommended. The processes of AI hallucination and fabrication must be addressed to protect patient safety, and the ethical concerns of utilizing AI within clinical settings must also be reconciled. Once these steps are taken, future iterations of ChatGPT and other forms of advanced AI may have the potential to be a clinical resource when assessing recommendations for various spinal conditions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.xnsj.2024.100333.

References

[1] Al Qaraghli MI, De Jesus O. Lumbar disc herniation. StatPearls. Treasure IslandFL: StatPearls Publishing; 2023.
[2] Alexander CE, Varacallo M. Lumbosacral radiculopathy. StatPearls. Treasure IslandFL: StatPearls Publishing; 2023.
[3] 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(1):180–91. doi:10.1016/j.spinee.2013.08.003.
[4] ChatGPT: optimizing language models for dialogue. OpenAI; 2022. Published November 30 Accessed October 1, 2023. Available from: https://openai.com/blog/chatgpt/ .
[5] 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(2):e0000198 Published 2023 Feb 9. doi:10.1371/journal.pdig.0000198.

[6] Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. Preprint. MedRxiv 2023:23285886 Published 2023 Feb 26. doi:10.1101/2023.02.21.23285886.
[7] Duey AH, Nietsch KS, Zaidat B, et al. Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations [published online ahead of print, 2023 Jul 25]. Spine J 2023;23(11):1684–91 S1529-9430(23)03285-0. doi:10.1016/j.spinee.2023.07.015.
[8] Rajjoub R, Arroyave JS, Zaidat B, et al. ChatGPT and its role in the decision-making for the diagnosis and treatment of lumbar spinal stenosis: a comparative analysis and narrative review [published online ahead of print, 2023 Aug 10]. Global Spine J 2023;14(3):21925682231195783. doi:10.1177/21925682231195783.
[9] Vogels EA. A majority of americans have heard of ChatGPT, but few have tried it themselves. Pew Research Center. Accessed October 8, 2023. https://www.pewresearch.org/short-reads/2023/05/24/a-majority-of-americans-have-heard-of-chatgpt-but-few-have-tried-it-themselves/
[10] Dhanvijay AKD, Pinjar MJ, Dhokane N, Sorte SR, Kumari A, Mondal H. Performance of large language models (ChatGPT, Bing Search, and Google Bard) in solving case vignettes in physiology. Cureus 2023;15(8):e42972 Published 2023 Aug 4. doi:10.7759/cureus.42972.
[11] Kumari A, Kumari A, Singh A, et al. Large language models in hematology case solving: a comparative study of ChatGPT-3.5, Google Bard, and Microsoft Bing. Cureus 2023;15(8):e43861 Published 2023 Aug 21. doi:10.7759/cureus.43861.
[12] Rosoł Maciej, Gąsior Jakub S, Łaba Jonasz, Korzeniewski Kacper, Młyńczak Marcel. Evaluation of the performance of GPT-3.5 and GPT-4 on the medical final examination. MedRxiv 2023:23290939. doi:10.1101/2023.06.04.23290939.
[13] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv 2023. doi:10.48550/arXiv.2303.13375.
[14] Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT–3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthopaed Surgeons 2023;31(23):1173–9 :10.5435/JAAOS-D-23-00396, September 4. doi:10.5435/JAAOS-D-23-00396.
[15] Chatterjee J, Dethlefs N. This new conversational AI model can be your friend, philosopher, and guide ... and even your worst enemy. Patterns (N Y) 2023;4(1):100676 Published 2023 Jan 13. doi:10.1016/j.patter.2022.100676.
[16] Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care 2023;27(1):120 Published 2023 Mar 21. doi:10.1186/s13054-023-04393-x.
[17] Fardon DF, Milette PC. Combined Task Forces of the North American Spine Society, American Society of Spine Radiology, and American Society of Neuroradiology. Nomenclature and classification of lumbar disc pathology. Recommendations of the Combined task Forces of the North American Spine Society, American Society of Spine Radiology, and American Society of Neuroradiology. Spine (Phila Pa 1976) 2001;26(5):E93–E113. doi:10.1097/00007632-200103010-00006.
[18] Jackson RP, Cain JE Jr, Jacobs RR, Cooper BR, McManus GE. The neuroradiographic diagnosis of lumbar herniated nucleus pulposus: II. A comparison of computed tomography (CT), myelography, CT-myelography, and magnetic resonance imaging. Spine (Phila Pa 1976) 1989;14(12):1362–7. doi:10.1097/00007632-198912000-00013.
[19] Pape E, Eldevik P, Vandvik B. Diagnostic validity of somatosensory evoked potentials in subgroups of patients with sciatica. Eur Spine J 2002;11(1):38–46. doi:10.1007/s005860100322.
[20] Vucetic N, Svensson O. Physical signs in lumbar disc hernia. Clin Orthop Relat Res 1996;333:192–201.
[21] Majlesi J, Togay H, Unalan H, Toprak S. The sensitivity and specificity of the Slump and the Straight Leg Raising tests in patients with lumbar disc herniation. J Clin Rheumatol 2008;14(2):87–91. doi:10.1097/RHU.0b013e31816b2f99.
[22] Bakhtiary AH, Safavi-Farokhi Z, Rezasoltani A. Lumbar stabilizing exercises improve activities of daily living in patients with lumbar disc herniation. J Back Musculoskelet Rehabilit 2005;18(3-4):55–60. doi:10.3233/bmr-2005-183-401.
[23] Genevay S, Viatte S, Finckh A, Zufferey P, Balagué F, Gabay C. Adalimumab in severe and acute sciatica: a multicenter, randomized, double-blind, placebo-controlled trial. Arthritis Rheum 2010;62(8):2339–46. doi:10.1002/art.27499.
[24] Clarke J, van Tulder M, Blomberg S, de Vet H, van der Heijden G, Bronfort G. Traction for low back pain with or without sciatica: an updated systematic review within the framework of the Cochrane collaboration. Spine (Phila Pa 1976) 2006;31(14):1591–9. doi:10.1097/01.brs.0000222043.09835.72.
[25] Ackerman WE AM. The efficacy of lumbar epidural steroid injections in patients with lumbar disc herniations. Anesth Analg 2007;104(5):1217–22. doi:10.1213/01.ane.0000260307.16555.7f.
[26] Peul WC, van Houwelingen HC, van den Hout WB, et al. Surgery versus prolonged conservative treatment for sciatica. N Engl J Med 2007;356(22):2245–56. doi:10.1056/NEJMoa064039.
[27] Kohlboeck G, Greimel KV, Piotrowski WP, et al. Prognosis of multifactorial outcome in lumbar discectomy: a prospective longitudinal study investigating patients with disc prolapse. Clin J Pain 2004;20(6):455–61. doi:10.1097/00002508-200411000-00011.
[28] Stroop A, Stroop T, Zawy Alsofy S, Nakamura M, Möllmann F, Greiner C, Stroop R. Large language models: Are artificial intelligence-based chatbots a reliable source of patient information for spinal surgery? Eur Spine J 2023. doi:10.1007/s00586-023-07975-z.
[29] Farhud DD, Zokaei S. Ethical issues of artificial intelligence in medicine and healthcare. Iran J Public Health 2021;50(11) i–v. doi:10.18502/ijph.v50i11.7600.
[30] Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018;378(11):981–3. doi:10.1056/NEJMp1714229.
[31] Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023;15(2):e35179 Published 2023 Feb 19. doi:10.7759/cureus.35179.
[32] Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;95:104770. doi:10.1016/j.ebiom.2023.104770.