Neurospine 2024;21(2):633-641.
https://doi.org/10.14245/ns.2448098.049

Neurospine
pISSN 2586-6583 eISSN 2586-6591

Original Article

Analyzing Large Language Models' Responses to Common Lumbar Spine Fusion Surgery Questions: A Comparison Between ChatGPT and Bard

Siegmund Philipp Lang[1,2], Ezra Tilahun Yoseph[1], Aneysis D. Gonzalez-Suarez[1], Robert Kim[1], Parastou Fatemi[3], Katherine Wagner[4], Nicolai Maldaner[1,5], Martin N. Stienen[6], Corinna Clio Zygourakis[1]

[1]Department of Neurosurgery, Stanford University School of Medicine, Stanford, CA, USA
[2]Department of Trauma Surgery, University Hospital Regensburg, Regensburg, Germany
[3]Department of Neurosurgery, Cleveland Clinic, Cleveland, OH, USA
[4]Ventura Neurosurgery, Ventura, CA, USA
[5]Department of Neurosurgery, University Hospital Zurich & Clinical Neuroscience Center, University of Zurich, Zurich, Switzerland
[6]Department of Neurosurgery & Spine Center of Eastern Switzerland, Cantonal Hospital St. Gallen & Medical School of St. Gallen, St. Gallen, Switzerland

Corresponding Author
Corinna Clio Zygourakis
https://orcid.org/0000-0002-1894-9252

Department of Neurosurgery, Stanford Neuroscience Health Center, Stanford University School of Medicine, 213 Quarry Rd, 2nd Floor, Stanford, CA 94304, USA
Email: corinnaz@stanford.edu

Received: January 1, 2024
Revised: March 19, 2024
Accepted: March 26, 2024

Objective: In the digital age, patients turn to online sources for lumbar spine fusion information, necessitating a careful study of large language models (LLMs) like chat generative pre-trained transformer (ChatGPT) for patient education.
Methods: Our study aims to assess the response quality of Open AI (artificial intelligence)'s ChatGPT 3.5 and Google's Bard to patient questions on lumbar spine fusion surgery. We identified 10 critical questions from 158 frequently asked ones via Google search, which were then presented to both chatbots. Five blinded spine surgeons rated the responses on a 4-point scale from 'unsatisfactory' to 'excellent.' The clarity and professionalism of the answers were also evaluated using a 5-point Likert scale.
Results: In our evaluation of 10 questions across ChatGPT 3.5 and Bard, 97% of responses were rated as excellent or satisfactory. Specifically, ChatGPT had 62% excellent and 32% minimally clarifying responses, with only 6% needing moderate or substantial clarification. Bard's responses were 66% excellent and 24% minimally clarifying, with 10% requiring more clarification. No significant difference was found in the overall rating distribution between the 2 models. Both struggled with 3 specific questions regarding surgical risks, success rates, and selection of surgical approaches (Q3, Q4, and Q5). Interrater reliability was low for both models (ChatGPT: k = 0.041, p = 0.622; Bard: k = -0.040, p = 0.601). While both scored well on understanding and empathy, Bard received marginally lower ratings in empathy and professionalism.
Conclusion: ChatGPT3.5 and Bard effectively answered lumbar spine fusion FAQs, but further training and research are needed to solidify LLMs' role in medical education and healthcare communication.

Keywords: Artificial intelligence, Large language models, Patient education, Lumbar spine fusion, ChatGPT, Bard

INTRODUCTION

Lumbar spine fusion surgery, a pivotal procedure in addressing diverse spinal pathologies, has evolved remarkably over recent years and is one of the most frequently performed neurosurgical procedures worldwide.[1] Due to multiple pathologies, presentations, and surgical approaches, patients may often find it daunting to understand the intricacies of lumbar fusion surgery, including its potential risks, benefits, and postoperative trajectories.[2]

In today's digital era, a substantial number of patients turn to online platforms for surgical information.[3] Encouraging patients to access treatment-related online health information can enhance patient compliance and the patient-physician relationship, while also enabling physicians to stay updated on emerging treatments.[3] However, the expansive digital domain occasionally presents conflicting, obsolete, or excessively technical data, potentially exacerbating patients' decision-making conundrum.[4] It is critical to understand that the quality of online information and its integration into medical consultations impacts patient care and patient-physician communication.[5] This highlights the need for accurate, accessible, and patient-centric online educational tools.

Artificial intelligence (AI) is transforming medicine in many ways, and the advent of large language models (LLMs) such as OpenAI's chat generative pre-trained transformer (ChatGPT) offers a transformative approach to patient education.[6] Leveraging their capability to sift through immense data and produce human-like narratives, LLMs can produce comprehensive, succinct, and individualized information to patients.[7] However, ChatGPT's real-world performance in complex fields like medicine and spine surgery still remains to be seen. Information obtained via LLMs may enable patients to better understand their disease process and treatment options, potentially improving the transparency and trust in the surgical decision-making process.[8] But there is also considerable risk to these models, including the potential for inaccurate or biased information that can mislead patients and negatively impact their health. The goal of this study is therefore to evaluate the accuracy, clarity, and comprehensiveness of Open AI's ChatGPT 3.5 and Google's Bard on 10 frequently asked patient questions regarding lumbar fusion surgery.

MATERIALS AND METHODS

A comprehensive Google search was conducted using the search terms "frequently asked questions AND lumbar spine surgery OR lumbar fusion surgery," yielding approximately 4,610,000 results within 0.51 seconds (September 7th, 2023; region: Germany). For this study, the first 20 Google hits were reviewed, and the following inclusion and exclusion criteria were applied (Table 1).

Concurrently, a research-specific search was executed on PubMed using the term "ChatGPT frequently asked patient questions." In addition, ChatGPT 4 was directly engaged with the prompt "Suggest a list of the 20 most common frequently asked patient questions about lumbar spine fusion surgery" prompting it to generate a list of questions relevant to our study.

This multiphased approach resulted in a consolidated pool of 158 questions, from which 10 frequently recurring topics emerged (Table 2, Supplementary Material 1). The authors reviewed this topic list and formulated 10 final questions that comprised the most critical and commonly addressed patient concerns on lumbar fusion surgery (Table 3). Fig. 1 presents a flowchart outlining this process to define 10 final questions.

The final questions were then submitted to the AI chatbot ChatGPT 3.5 through its online portal (https://chat.openai.com/chat) on October 21, 2023, using the following prompt (Answer Set #1): "*Act as an expert spine surgeon who is up to date with the latest scientific research and has years of experience counseling patients with empathy and clarity. Provide a comprehensive and easily understandable answer to the following question about lumbar spine fusion surgery. Limit your answer to 150 words and focus on the most important aspects*." The same questions and prompt were also presented to Google's Bard (https://bard.google.com/chat) on the same date (Answer Set #2). For each question, a

Table 1. Inclusion and exclusion criteria for questions

Inclusion criteria	Exclusion criteria
Published after January 1st 2017	Nongeneralizable information, e.g., provider or implant specific details
Published in English language	Emphasis on nonneurosurgical aspects, e.g., anesthesiological information
Information presented in FAQ or Q&A sections	

FAQ, frequently asked questions; Q&A, questions and answers.

https://doi.org/10.14245/ns.2448098.049

Table 2. Ten most frequent topics

Ranking	Topic
1	Definition and purpose (e.g., "What is spine fusion (surgery)?," "Why is spine fusion done?," etc.) - The intention behind these questions is often to understand the nature and objective of the procedure. Frequency: 10 times
2	Duration & recovery (e.g., "How long does spine fusion surgery take?," "How long is the recovery after surgery?") - This theme relates to the time-related aspects of the surgery and the recovery phase. Frequency: 9 times
3	Risks & complications (e.g., "What are the risks of spine fusion surgery?," "How common is infection after surgery?," etc.) - These questions highlight potential adverse outcomes or challenges following the procedure. Frequency: 8 times
4	Success rate & outcome (e.g., "What is the success rate of spine fusion surgery?") - The goal here is to gauge the probable effectiveness and positive results of the surgery. Frequency: 7 times
5	Approach/method (e.g., "Which approach is better for spine fusion surgery? Anterior or Posterior?") - These questions aim to understand the techniques and strategies involved in the procedure. Frequency: 6 times
6	Postsurgery limitations & lifestyle (e.g., "What limitations will I have after spine fusion surgery?," "Can I do any activity I want after the surgery?") - These questions seek clarity on how life may change or be restricted following the surgery. Frequency: 5 times
7	Preparation & criteria (e.g., "When should I get spine fusion surgery?," "Is outpatient spine fusion surgery safe?") - This theme encompasses questions related to deciding on the surgery and understanding preparatory requirements. Frequency: 5 times
8	Postsurgery care & maintenance (e.g., "How should I care for my back after spine fusion surgery?," "What physiotherapy or exercises are recommended postsurgery?") - These questions cover the care and activities required after the surgery to ensure recovery and maintain health. Frequency: 4 times
9	Comparisons with other procedures (e.g., "How does spine fusion surgery compare with disc replacement?," "Is spine fusion better than laminectomy?") - These questions look to understand the procedure in relation to other similar or alternative treatments. Frequency: 4 times
10	Costs & insurance (e.g., "How much does spine fusion surgery typically cost?," "Is the surgery covered by insurance?") - Financial aspects and concerns about the procedure are addressed in this theme. Frequency: 2 times

new window was created in the respective Chatbot to avoid bias from the prior questions. A list of all answers provided by the 2 Chatbots can be found in Supplementary Material 2. ChatGPT, by OpenAI, utilizes the generative pre-trained transformer (GPT) architecture, focusing on generating broad, versatile responses across various topics through deep learning techniques.[9] The model is pretrained on a diverse range of internet texts, allowing it to generate responses across a wide array of topics. Its iterative training and updates, such as the transition from GPT 3 to GPT 3.5 and beyond, focus on improving its understanding, accuracy, and ability to generate human-like text based on the input it receives. Google's Bard, on the other hand, leverages language model for dialogue applications (LaMDA), a system designed specifically to handle conversational applications.[10] LaMDA's training regime includes a blend of reinforcement learning from human feedback and other methods to fine-tune its performance in dialogue-based tasks. This focus aims to produce more relevant and contextually appropriate responses,

https://doi.org/10.14245/ns.2448098.049

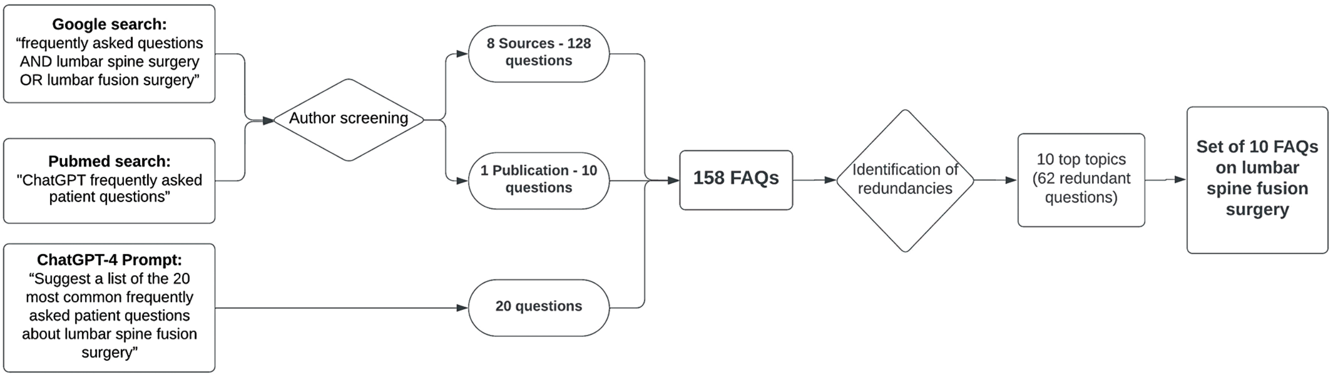
Google search:
"frequently asked questions AND lumbar spine surgery OR lumbar fusion surgery"
Pubmed search:
"ChatGPT frequently asked patient questions"
ChatGPT-4 Prompt:
"Suggest a list of the 20 most common frequently asked patient questions about lumbar spine fusion surgery"
Author screening
8 Sources - 128 questions
1 Publication - 10 questions
20 questions
158 FAQs
Identification of redundancies
10 top topics (62 redundant questions)
Set of 10 FAQs on lumbar spine fusion surgery

Fig. 1. Selection process for identifying top 10 FAQs on lumbar spine fusion surgery. This flowchart illustrates the methodology from initial search to final question curation. ChatGPT, chat generative pre-trained transformer; FAQ, frequently asked questions.

Table 3. Ten FAQs that were presented to the large language models

No.	FAQs
1	What is lumbar spine fusion surgery?
2	How long is the recovery after lumbar spine fusion surgery?
3	What are the potential risks and complications associated with lumbar spine fusion surgery?
4	What is the success rate of lumbar spine fusion surgery?
5	Which surgical approach is the best for lumbar spine fusion?
6	What limitations should I expect after undergoing lumbar spine fusion surgery?
7	How long is the hospital stay after lumbar spine fusion surgery?
8	How should I care for my back after lumbar spine fusion surgery to ensure the best recovery?
9	What are the alternatives to lumbar spine fusion surgery?
10	Is lumbar spine fusion surgery typically covered by insurance?

FAQ, frequently asked questions.

Table 4. Supplementary evaluation criteria for each data set

No.	Evaluation criteria
1	The overall content of all answers is comprehensive and covers all necessary aspects.
2	The answers are easy to understand and are communicated clearly.
3	The answers address patient concerns empathetically and professionally.
4	The overall length and detail of each answer are appropriate for the target audience.

especially in conversational settings. This technical divergence—ChatGPT's wide-ranging generative capabilities versus Bard's conversational precision—highlights their potential differences in applicability to medical education and patient communication.

Five blinded spine surgeons (either in spine-fellowship or spine-fellowship trained attending spine surgeons who did not know that these were LLM-generated responses) rated each response using a previously published rating system[11]: 'excellent response not requiring clarification,' 'satisfactory requiring minimal clarification,' 'satisfactory requiring moderate clarification,' or 'unsatisfactory requiring substantial clarification.' Satisfactory responses conveyed primarily factual data, largely devoid of inaccuracies, albeit necessitating some elucidation. Responses warranting 'minimal clarification' were factually accurate but either lacked comprehensive information or failed to capture nuances from the literature. Those necessitating 'moderate clarification' relayed obsolete or irrelevant data. A response was deemed unsatisfactory if it encompassed data that was either outdated or overly generic, rendering it susceptible to misinterpretation.[11]

The blinded raters also responded to 2 additional questions (Table 4) on a 5-point Likert scale (from "I strongly disagree" to "I strongly agree") to assess whether the responses were easy to understand and clearly communicated, as well as whether they address patient concerns empathetically and with professionalism.

Data are presented using absolute values, percentages, mean and standard deviations for descriptive purposes. The word counts from ChatGPT 3.5, and Bard were compared using an independent t-test. To assess the relationship between word count and median ratings for each model, Pearson correlation coefficient (r) was calculated. The distribution of answer ratings across predefined categories was analyzed using the Wilcoxon signed-rank test, where 'W' is the test statistic. This test was chosen to compare the differences in ratings between ChatGPT

https://doi.org/10.14245/ns.2448098.049

3.5 and Bard. The Mann-Whitney U-test was used to compare the ratings between the 2 models for each of the 10 questions. The interrater reliability for the ratings was assessed using Cohen kappa. All statistical procedures were performed using IBM SPSS Statistics ver. 28.0 (IBM Co., Armonk, NY, USA) and Excel version 2302 (Microsoft 365, Microsoft, Redmond, WA, USA). The level of statistical significance was set at $p < 0.05$.

This study was exempt from Institutional Review Board review.

RESULTS

Despite being prompted to limit the answer to 150 words each, Bard's answers had a significantly higher word count of 202.6 ± 42.9 (range, 138–287) compared to ChatGPT 3.5 (mean, 158.9 ± 18.1; range, 127–189; $p < 0.05$). ChatGPT's word count was positively correlated with the median rating ($r = 0.735$, $p < 0.05$), while Bard's word count was negatively correlated with the median rating ($r = -0.68$, $p < 0.05$).

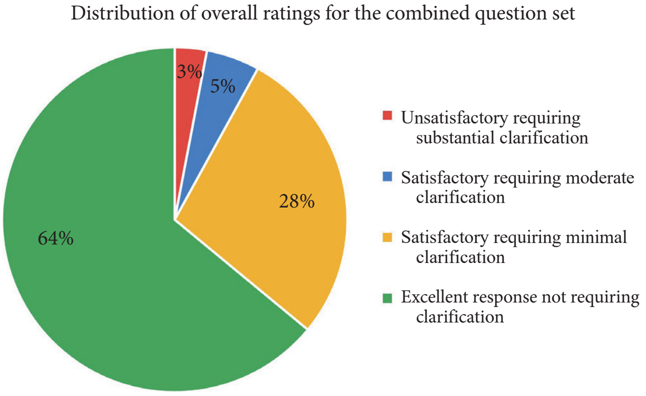

Fig. 2. Pie chart with the distribution of overall ratings, expressed in percentages, for the combined question set across the 2 large language models.

Across both models and all 10 questions, 97% of answers were rated as satisfactory, and only 3% were unsatisfactory. Specifically, 64% (n = 64) were excellent without any clarification needed, 28% (n = 28) were satisfactory requiring minimal clarification, 5% (n = 5) were satisfactory requiring moderate clarification, and 3% (n = 3) were unsatisfactory requiring substantial clarification (Fig. 2).

For ChatGPT across all 10 questions, 62% of responses were excellent without any clarification needed, 32% were satisfactory requiring minimal clarification, 2% were satisfactory requiring moderate clarification, and 4% were unsatisfactory requiring substantial clarification. Bard had a slightly, but not statistically significantly, better performance, with only 2% of responses being rated as unsatisfactory requiring substantial clarification. For Bard, 66% of responses were excellent, 24% were satisfactory requiring minimal clarification, and 8% were satisfactory requiring moderate clarification. There was no statistically significant difference in the overall distribution of ratings between the 2 answer sets (ChatGPT 3.5 vs. Bard, $W = 12$; $p = 1$).

For the ChatGPT model, 7 out of 10 questions received median ratings of "excellent". Questions Q3, Q4, and Q5 had the lowest median ratings of "satisfactory requiring minimal clarification" (Fig. 3). The lowest median ratings were seen in Q3 and Q5 for the Bard responses, both of which had median respons-

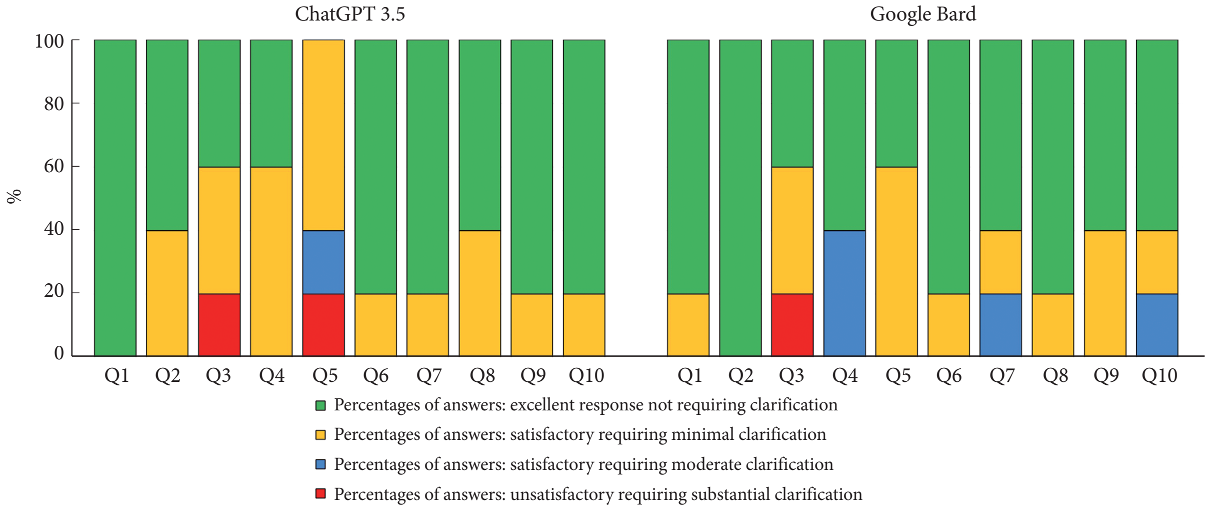

Fig. 3. Distribution of rater evaluations for ChatGPT and Bard across 10 lumbar surgery questions. The bars represent the percentage of raters assigning each category. ChatGPT, chat generative pre-trained transformer.

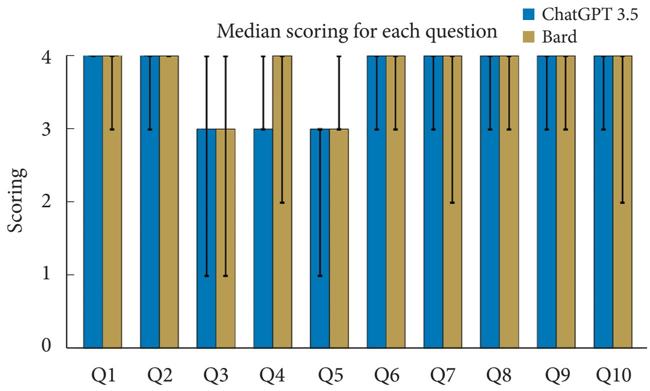

Fig. 4. Median ratings comparison between ChatGPT 3.5 and Bard. Bars show median scores; error bars show range (minimum–maximum scoring). No significant differences in ratings across all questions. ChatGPT, chat generative pretrained transformer.

es of "satisfactory requiring minimal clarification" (Fig. 3).

There was no statistically significant difference in the ratings between the 10 answers from ChatGPT compared to Bard (Fig. 4; Q1: p = 0.42, Q2: p = 0.18, Q3: p = 1.0, Q4: p = 1.0, Q5: p = 0.08, Q6: p = 1.0, Q7: p = 0.52, Q8: p = 0.60, Q9: p = 0.60, and Q10: p = 0.52). Both model responses had poor interrater reliability (ChatGPT: k = 0.041, p = 0.622; Bard: k = -0.040, p = 0.601).

For the questions assessing clarity/easiness to understand and empathy/professionalism, the median ratings for both ChatGPT and Bard were 5/5. This indicates that on average, raters found the answers from both models to be easy to understand, clearly communicated, and addressing patient concerns empathetically and professionally. There was no statistically significant difference in the median scores between ChatGPT and Bard for these 2 questions. However, a select few individual raters rated Bard slightly lower on empathy and professionalism in addressing patient concerns.

DISCUSSION

AI, particularly language models like ChatGPT, is increasingly being explored for its potential in patient education. Recent studies have begun to elucidate the capabilities and limitations of AI in different domains of medicine with a special focus on patient education.[11,12] In our current study, we systematically identified common patient questions on lumbar spine fusion surgery and evaluated the answers given by ChatGPT 3.5 and Google Bard based on ratings by "blinded" spine-fellowship trained surgeons.

ChatGPT adhered more closely to the specified 150-word limit than Bard. Notably, a positive correlation existed between ChatGPT's word count and higher median ratings, whereas Bard's word count showed a negative correlation with high ratings. This suggests that ChatGPT's longer responses were received more favorably, indicating that its additional content was perceived as valuable. Conversely, Bard's longer answers were seen as less effective or relevant, negatively impacting their reception.

The majority of responses for both models were excellent and did not require any clarification. Bard slightly outperformed ChatGPT in the proportion of excellent ratings, although this difference was not statistically significant. In addition, and very importantly, both models also achieved very high scores in empathy and professionalism. The slightly lower individual ratings for Bard on empathy and professionalism could imply a perceived difference in tone or response style.

Our results are consistent with findings from Ayers et al.,[12] who reported very high quality and empathy of AI-generated responses. In fact, in their cross-sectional study of 195 randomly drawn patient questions from a social media forum, they found that AI-generated responses had higher quality and empathy than physician-generated ones. This suggests that in the future, AI may be used to draft initial responses to patient queries, which can then be reviewed and personalized by clinicians. Such an approach could revolutionize patient-physician interactions and improve efficiency. In our current study, we did not compare the answers given by the LLMs with answers by physicians but focused on assessing the accuracy and reliability of the LLM-provided answers. The continual alignment of AI-generated information with expert knowledge, especially in the sensitive field of medicine, will pose an ongoing challenge for research in the coming years.

Questions Q3 (risks and complications of surgery), Q4 (surgical success rate), and Q5 (approach/method) for both models received lower median ratings, falling into the 'satisfactory requiring minimal clarification' category. This suggests specific areas where both models need to be improved and fine-tuned. These questions represent areas where the application of AI in medical contexts faces limitations due to the need for highly individualized, context-sensitive, and up-to-date information, which can be challenging for an AI model to synthesize and communicate effectively.

1. Surgical Risks and Complications

Question 3 focuses on the risks and complications of lumbar fusion surgery. The complexity in answering this question lies

https://doi.org/10.14245/ns.2448098.049

in the multifaceted nature of surgical risks and complications. Lumbar spine fusion surgery involves numerous variables, including patient-specific factors (age, health status, underlying conditions), surgical techniques, and postoperative care. AI models may struggle to integrate and personalize this vast, variable data into a concise, patient-specific response. This was summarized in ChatGPT's answer: "…While these risks exist, they are relatively rare, and we strive to minimize them. Your individual health, surgical technique, and postopetative care all play a role in the outcome. We prioritize your well-being and aim to provide the best possible results." Similarly, Bard summarized: "…The risks of lumbar spine fusion surgery vary depending on several factors, such as the patient's age, overall health, and the extent of the surgery. Your surgeon will discuss the risks and benefits of surgery with you in detail before you make a decision about whether or not to proceed." Indeed, the challenges faced by AI models like ChatGPT and Bard in addressing questions about risks and complications stem from the inherent multidimensionality of the topic.

2. Surgical Success Rate

Question 4 regarding surgical success rate was another question with overall lower ratings. The success rate of lumbar spine fusion surgery can be challenging to quantify due to varying definitions of "success," such as pain relief, functional improvement, spinal stability, appropriate hardware placement, and/or spinal bony fusion. These outcomes can differ significantly based on individual patient characteristics, the specific pathology being treated, and the surgical techniques used. It is also important to consider from whose perspective we are determining the success of the surgery, as this could be different for the surgeon versus the patient themselves versus the patient's caregiver. AI models may find it difficult to provide a generalized answer that accurately reflects the nuanced outcomes of different patient scenarios. This complexity is acknowledged in ChatGPTs answer for Q4: "Lumbar spine fusion surgery can be highly successful, but success rates vary depending on individual factors and the specific condition being treated. On average, the success rate falls between 60% to 90%. Factors influencing success include patient age, overall health, the surgeon's skill, and the reason for surgery…". Bard provided a similar answer: "The success rate of lumbar spine fusion surgery is generally considered to be high, with studies reporting success rates ranging from 70% to 90%. Success is typically defined as a significant reduction in pain and improvement in function. The success rate of lumbar spine fusion surgery can vary depending on a number of factors, such as the underlying condition being treated, the surgical approach used, the skill of the surgeon, and the patient's overall health…".

Both ChatGPT's and Bard's responses accurately capture the spectrum of factors that can influence the success of a patient's outcomes from spine fusion surgery.

3. Surgical Approach

Question 5 on the best surgical approach is inherently complex due to the evolving nature of surgical techniques and the lack of a universally accepted 'best' approach. Factors influencing the choice of surgical approach include the specific spinal pathology, surgeon's expertise, and technological advancements. AI models may struggle to make definitive recommendations without considering the individualized context of each patient's case. ChatGPT's answer for this question was: "The 'best' approach depends on your diagnosis, overall health, and surgeon's expertise. It's vital to discuss these options with your surgeon, who will consider your specific case and recommend the most appropriate approach." Bard listed posterior lumbar interbody fusion, anterior lumbar interbody fusion, and transforaminal lumbar interbody fusion and pointed out the growing trend of minimally invasive spine surgery. It went on to say: "Again, it has been underlined, that the best surgical approach for lumbar spine fusion depends on a number of factors, including the specific condition being treated, the number of levels being fused, and the patient's anatomy…".

Neither ChatGPT nor Bard included lateral lumbar interbody fusion (LLIF) in their responses. LLIF has emerged as a viable surgical approach, particularly noted for its less invasive nature and reduced disruption of posterior spinal elements. Studies have shown LLIF to be effective in achieving spinal fusion with potentially lower complication rates and less postoperative pain, especially beneficial in multilevel spine disorders and deformities.[13]

Both AI models' responses, while broad and general, aptly capture the multifactorial and individualized nature of determining the 'best' surgical approach. Studies have shown that the complication rates of a surgeon's surgical approach depend on several factors, including the patient's unique pathology, the number of spine levels involved, or the specific surgical devices and techniques employed—all of which shape the risk and outcome profile of the surgical intervention.[14,15] This underscores the need for a personalized approach in surgical planning to optimize patient outcomes.

Our study is limited in that it is a small study with only 5 raters. We had poor interrater reliability, suggesting that there may be differences in subjective interpretation of the models' answers, and/or different expectations for what constitutes a clear and comprehensive answer. In addition, our raters were all practicing spine surgeons, and we did not include any patients themselves. As a result, our findings do not necessarily reflect the patient perspective on the clarity and utility of the Chatbot responses to patient questions on lumbar fusion surgery. In addition, newer versions of LLMs have already been developed since this analysis was performed, and will continue to evolve at a rapid rate, potentially further improving the AI-generated responses. It's important to note our study intentionally focused on evaluating LLMs based on real-world, patient-posed questions from frequently asked question (FAQ) sections, rather than optimizing inquiries for maximal LLM performance. This approach aims to provide insights into the actual advice patients might encounter online, acknowledging a potential trade-off in analytical precision. Further, our study did not assess the consistency of LLM responses to the same question asked multiple times. This decision was based on our aim to evaluate the LLMs' performance in typical, real-life single-query interactions rather than exploring the variability of responses.

Recognizing that not all individuals with spinal conditions may be proficient in using AI, our study specifically evaluates LLM efficacy for users who engage with these platforms. This focus aims to shed light on AI's capabilities and constraints in enriching patient education among digitally inclined segments of the population.

In the future, incorporating individualized patient data, as well as data from specific surgeons and spine centers, into AI-based LLMs could significantly enhance the precision and relevance of the information provided by AI, making it a more effective tool in patient education and decision-making. Personalized data would allow for more tailored responses regarding risks and outcomes, while center-specific data can inform patients about the practices and success rates of particular surgeons or facilities. This approach could not only aid in informed decision-making but also facilitates quality improvement and benchmarking in medical practices.[16] However, ensuring data privacy and ethical use of this data will be essential in this process.

A recent study by Rajjoub et al.[17] assessed ChatGPT's responses against the 2011 North American Spine Society Clinical Guideline for lumbar spinal stenosis (LSS). This comparative analysis revealed that ChatGPT's responses were congruent with the current literature on LSS. Specifically, the study found alignment in ChatGPT's answers regarding the definition, diagnostic tests, and both nonsurgical and surgical interventions for LSS. The authors suggested that ChatGPT can effectively support the decision-making process for LSS diagnosis and treatment, potentially making it a valuable tool in the context of lumbar spine fusion surgery education.

Beyond specific surgical contexts, AI's role in patient education spans various areas. An article from the American Medical Association's Journal of Ethics highlights the potential for AI to enhance patient-clinician relationships.[18] The authors suggested, that by automating routine inquiries and administrative tasks, AI could allow clinicians to focus more on patient interaction and relationship-building. This aspect is particularly relevant for patient education, where AI could provide detailed and personalized information about treatment options, thus facilitating shared decision-making.

CONCLUSION

LLMs like ChatGPT and Bard hold significant promise for patient education in lumbar spine fusion surgery and broader medical contexts. In this study, we find that both LLMs produce accurate, clear, and empathetic responses to the most commonly asked questions about spinal fusion surgery. Nevertheless, human oversight remains crucial to ensure the effective and appropriate use of AI in healthcare. Training the models with patient-, surgeon-, and center-specific data may potentially increase their value. Future research should continue to explore and refine AI's role, aiming for a harmonious integration of technology and human expertise in patient care and education.

NOTES

Supplementary Material: Supplementary materials 1-2 can be found via https://doi.org/10.14245/ns.2448098.049.

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: SPL, CCZ; Formal analysis: SPL, ETY, ADGS, RK, PF, KW, NM, MNS; Methodology: SPL, ETY, ADGS, RK, PF, KW, NM, MNS, CCZ; Project administration: SPL, CCZ; Writing – original draft: SPL, ETY, ADGS, CCZ; Writing – review & editing: SPL, ETY, ADGS, RK, PF, KW, NM, MNS, CCZ.

ORCID

Siegmund Philipp Lang: 0000-0003-0459-9092
Ezra Tilahun Yoseph: 0000-0002-4184-9292
Aneysis D. Gonzalez-Suarez: 0000-0003-3745-763X
Parastou Fatemi: 0000-0001-8188-8440
Katherine Wagner: 0000-0002-5976-828X
Nicolai Maldaner: 0000-0003-0284-1033
Martin N. Stienen: 0000-0002-6417-1787
Corinna Clio Zygourakis: 0000-0002-1894-9252

REFERENCES

1. Mobbs RJ, Phan K, Malham G, et al. Lumbar interbody fusion: techniques, indications and comparison of interbody fusion options including PLIF, TLIF, MI-TLIF, OLIF/ATP, LLIF and ALIF. J Spine Surg 2015;1:2-18.
2. Gaudin D, Krafcik BM, Mansour TR, et al. Considerations in spinal fusion surgery for chronic lumbar pain: psychosocial factors, rating scales, and perioperative patient education-a review of the literature. World Neurosurg 2017;98: 21-7.
3. Zhang Z, Yang H, He J, et al. The Impact of treatment-related internet health information seeking on patient compliance. Telemed J E Health 2021;27:513-24.
4. Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Educ Res 2001; 16:671-92.
5. Langford AT, Roberts T, Gupta J, et al. Impact of the internet on patient-physician communication. Eur Urol Focus 2020; 6:440-4.
6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med 2023;29:1930-40.
7. Hung YC, Chaker SC, Sigel M, et al. Comparison of patient education materials generated by chat generative pre-trained transformer versus experts: an innovative way to increase readability of patient education materials. Ann Plast Surg 2023;91:409-12.
8. Blease C, Bernstein MH, Gaab J, et al. Computerization and the future of primary care: a survey of general practitioners in the UK. PloS One 2018;13:e0207418.
9. Yenduri G, M R, G CS, et al. Generative pre-trained transformer: a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. arXiv:2305.10435v2 [Preprint]. 2023 [cited 2024 Mar 5]. Available from: http://arxiv.org/abs/2305.10435.
10. Thoppilan R, De Freitas D, Hall J, et al. LaMDA: language models for dialog applications. arXiv:2201.08239v3 [Preprint]. 2022 [cited 2024 Mar 5]. Available from: http://arxiv.org/abs/2201.08239.
11. Mika AP, Martin JR, Engstrom SM, et al. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. J Bone Joint Surg Am 2023;105:1519-26.
12. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183:589-96.
13. Hijji FY, Narain AS, Bohl DD, et al. Lateral lumbar interbody fusion: a systematic review of complication rates. Spine J 2017;17:1412-9.
14. Lee MJ, Konodi MA, Cizik AM, et al. Risk factors for medical complication after spine surgery: a multivariate analysis of 1,591 patients. Spine J 2012;12:197-206.
15. Lange N, Stadtmüller T, Scheibel S, et al. Analysis of risk factors for perioperative complications in spine surgery. Sci Rep 2022;12:14350.
16. Johnson KB, Wei W, Weeraratne D, et al. Precision medicine, ai, and the future of personalized health care. Clin Transl Sci 2021;14:86-93.
17. 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. Global Spine J 2024;14:998-1017.
18. Nagy M, Sisk B. How will artificial intelligence affect patient-clinician relationships? AMA J Ethics 2020;22:E395-400.

https://doi.org/10.14245/ns.2448098.049