Title: AI vs. Surgeons: A Comparative Analysis of Perspectives on Controversial Spinal Pathology Management

Student number: 2001084

Name: Saylan Mehmet

Design: Research paper

Institution: UHW

Tutor: Michael McCarthy

Ethical approval: Not required

AI vs. Surgeons: A Comparative Analysis of Perspectives on Controversial Spinal Pathology Management

Saylan Mehmet

Objective: To assess the capabilities of LLMs to provide definitive and accurate treatment plans for controversially managed spinal surgery patients.
Design: Research paper
Setting: University Hospital Wales
Participants: 8 Consultant spinal surgeons and 3 Large Language Models (LLMs): ChatGPT3, ChatGPT4 and Bard.
Results: There was a low agreement rate between the LLMs and surgeons overall, with ChatGPT4 being too cautious to answer definitively/ at all on multiple occasions, and Bard performing over-confidently frequently generating incorrect, and unsafe responses.
Conclusions: LLMs are not currently suitable to be relied upon for managing controversially managed spinal conditions, however, could be used as an educational tool with user caution of "hallucinations".

INTRODUCTION

Spinal pathologies are complex and often require the expertise of highly trained spinal surgeons for appropriate diagnosis and management. With advances in technology, artificial intelligence (AI) has been increasingly utilized in the field of spinal surgery, including for image analysis, surgical planning, and intraoperative guidance[1,2]. AI has shown promising results in assisting surgeons in decision-making and improving surgical outcomes[3].

The use of AI in spinal surgery raises questions about its potential to replace or supplement the skills and knowledge of human spinal surgeons. Some studies have shown that AI can outperform human clinicians in diagnostic accuracy and decision-making[4,5]. However, concerns have also been raised about the limitations of AI, including the lack of contextual knowledge, inability to perform physical examination, and potential biases in the algorithms[6,7].

Large Language Models (LLMs) are a type of AI that have significantly advanced in recent times and proven to be extremely capable. These models are publicly accessible as online chatbots in which users can enter substantial amounts of information and requests to generate contextually relevant answers and solutions. These LLMs use deep learning techniques and natural language processing (NLP) whilst trained on a large corpus of data in order to produce human-like responses[8]. Currently, the most popular LLM-powered Chatbots are ChatGPT (Generative Pretrained Transformer), Google Bard, and Bing Chat.

LLMs are now being applied to solve a variety of problems with subjective answers within medicine, having shown exceedingly promising results across several fields already. ChatGPT has been used to take the United States Medical Licencing Exam (USMLE), consisting of 3 exams, performing at or near the passing threshold for all 3 without any specialised

training or reinforcement[9]. Within radiology, it can simplify radiology reports with most generated being factually correct, complete, and not potentially harmful to the patient[10]. Additionally, it can provide antimicrobial advice on open-ended medical queries almost as well as the average human physician could[11].

ChatGPT is an NLP model developed and launched by OpenAI in November 2022, following their first ever LLMs (GPT and GPT2) developed in 2018-2019. It was designed to generate human-like responses to be used in chatbot conversations[12]. ChatGPT's training dataset, whilst ever-growing, is on 300 billion words, (570 GB) from crawled web, books, Wikipedia, and many other sources. Since its launch, OpenAI have released multiple versions of ChatGPT, the most recent being ChatGPT4, with its predecessors being ChatGPT3.5 Default and ChatGPT3.5 Legacy.

Another popular LLM is Google's Bard, released in March 2023 as a more lightweight and optimised version of LaMDA (Language Model for Dialogue Applications), another LLM, intended to rival the capabilities of ChatGPT. Like ChatGPT, Google Bard was also trained on a massive dataset, but of text and code, containing 1.56 trillion words, however unlike ChatGPT, the dataset that Bard uses is current, whereas ChatGPT's dataset is only up until 2021. This dataset is a combination of text from the public domain, such as books, articles, and code, as well as text from Google's internal systems.

Lastly, Bing Chat, uses an improved version of the ChatGPT model called Prometheus, and was designed to produce refined results, comprehensive answers, and creative inspiration, acting as an AI-powered "co-pilot" for a user's internet searches.

With this information in mind, and AI's ever-growing day-to day use, we aimed to apply this in medicine and answer whether the vast datasets that these LLM-powered chatbots were trained on meant that they had the ability to match the performance, or even outperform 8 consultant spinal surgeons with a combined 153 years of experience in spinal surgery. The LLMs that were analysed and compared to the surgeons were ChatGPT3.5 (default), ChatGPT4, Google Bard and Bing Chat.

OBJECTIVES

To date, there have been few studies directly comparing the performance of AI and human spinal surgeons in diagnosing and managing complex spinal scenarios. This paper aims to fill this gap by comparing the diagnostic accuracy and treatment outcomes of AI and spinal surgeons in a series of 18 complex spinal cases. By examining the strengths and limitations of each approach, we aim to provide insight into the potential role of AI in spinal surgery, decision making, its safety, accuracy, and its impact on the work of spinal surgeons as well as its future potentials in the field.

METHODS

The senior author devised 18 spinal scenarios with a total of 55 questions based on known common controversial spinal scenarios and review of literature and textbooks. The senior author asked 8 consultants if they were interested in participating in the AI vs Surgeons project, with a response rate of 100%. They were each presented with a document of compiled questions with options to choose from and space to explain the reasoning behind each choice of answer. This was to mimic the response style of the LLMs which provided an answer as well as its reasoning. This document was completed by each consultant twice, leaving a couple of weeks between responses to assess the reproducibility of the answers.

The same questions were then asked to the LLMs, with each question's response being regenerated 5 times for reproducibility reasons once more. The first LLM asked was ChatGPT3.5, then ChatGPT4, then Google Bard, and lastly Bing Chat. Bing Chat was excluded from the study due to the reluctance to answer any of the questions, with each response stating to seek professional medical advice instead, thus leaving the 3 LLMs. Each response was documented on a spreadsheet for later analysis.

For each of the questions the agreement, most chosen answer, and the number of times the LLM abstained from providing an answer was calculated for each of the 5 responses for every question. This enabled a consensus option from the most chosen responses across the 3 LLMs to be reached for each question. Each consensus option was then compared to the surgeon's responses to assess whether the surgeon's agreed with the LLMs.

Whilst generating the responses from the LLMs, two themes became apparent: the occasional indecisiveness of the responses, as well as the strong encouragement to ask a professional doctor, likely a safety net feature. Whether each LLM was decisive or not for each response was noted, as well as the overall percentage of decisiveness for each response. The same was done based on the LLMs suggestion to ask a doctor.

Since no real patient information or features were identified, nor were clinical pictures used, ethical approval for this project was not required.

The 15 responses from each question presented to the 3 LLMs were then presented to ChatGPT3.5 to identify the themes within the responses, and to rank the responses based on how they each fulfil these themes. This thematic analysis was done for each question.

For statistical analysis, Statistical Package for the Social Sciences (SPSS) was used. This required inputting our spreadsheet of data collected from the surgeons and LLMs and cleaning it to ensure all missing values, errors or outliers were corrected. Once the statistics were obtained, they were exported to other software programs, allowing the data to be made into visual representations and into a report.

RESULTS AND DISCUSSION

By analysing the frequency of which the LLMs suggested to ask a healthcare professional in response to a question, whether they successfully answered the question or not, can tell us many things. Figure 1 shows a breakdown of the number of times this was done.

AI	Number of times suggesting to ask a doctor	%
ChatGPT3.5	168	61
ChatGPT4	272	99
Bard	123	45

Figure 1: The number of times out of the 275 responses that the LLM's advised to seek professional medical help.

In total, with 5 response regenerations per question and 55 questions, this equates to a total number of 275 responses per LLM. The table in Figure 1 shows that ChatGPT4 advised the user to seek professional medical help from a doctor 272 out of 275 times, 99% of the time, whereas Bard made this suggestion 123 out of 275 times, 45% of the time. This indicates a few things. Firstly, the large increase in the percentage of times this suggestion was made by ChatGPT3.5 and ChatGPT4 (an increase by 62.3%) highlights the revisions done to ChatGPT3.5 prior to releasing the newer version. These revisions being not only in terms of increasing the abilities of what the LLM can do, for example, ChatGPT3.5 is a text-input only model, however ChatGPT4 allows input of other forms of data such as images, but other revisions made include those regarding safety. It is clear from these results that the nearly 100% of times ChatGPT4 advised the user to seek professional help was due to a safety net feature added following OpenAI's discovery of ChatGPT3.5's frequent use in medical research. It is no doubt that for general members of the public who might be using ChatGPT to research their medical queries, that the chatbot may potentially be harmful in misdiagnosing them. Thus, whilst ChatGPT is happy to provide suggestions following medical queries, for patient safety, the updated model, ChatGPT4 makes it very clear that it is only AI which is inferior to the knowledge of a real healthcare professional. This avoids patients treating the LLM as a supplementation to real life healthcare advice from a doctor.

Secondly, for the LLMs displaying a lower percentage of times they suggested to ask a doctor, this can be implied as 'confidence' by the user, which again, could be dangerous. Bard particularly had this problem. Bard would frequently respond to the controversial questions quite confidently, giving one concise answer with no suggestion to seek further advice from a real doctor, as if the question was a simple one, when in fact, these responses were providing incorrect medical advice multiple times. Since these LLMs have been trained in NLP and so mimic human responses to the extent of having tones of emotion in their responses, Bard in this sense compared to ChatGPT had "cocky" responses, confidently answering questions incorrectly.

In the AI world, this is what is termed "hallucination" in which the LLM gives a confident and convincing but made-up answer due to errors in encoding and decoding[13]. These hallucinations can be extensive and incredibly realistic as shown in an example, Figure 2, that I made up and is completely non-existent. This scarily displays how hallucinations in medical advice could harm patients with this misleading confident tone. Not only were the hallucinations convincing in tone, but when asked to provide references for each response, occasionally, ChatGPT would provide references which had absolutely nothing to do with the response, and worse, occasionally could not provide any references at all. ChatGPT's hallucination rate is approximately 15-20%[14], making it one of OpenAI's biggest challenges to continue to decrease this percentages with subsequent updated models.

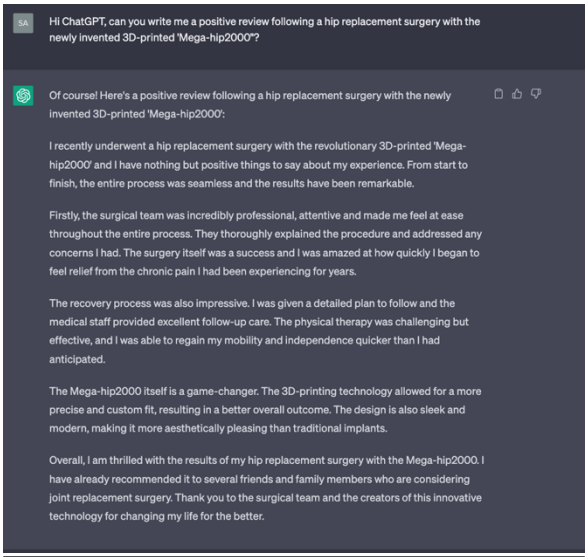

Figure 2: Example of a ChatGPT "hallucination".

Whilst the point of this research is to see how LLMs cope with answering controversial questions, some LLMs gave more definitive answers than others shown in Figure 3 below.

AI	Number of decisive answers	%
ChatGPT3.5	196	71
ChatGPT4	65	24
Bard	252	92

Figure 3: The number of decisive, definitive answers the LLMs gave out of 275 responses.

The data above is consistent with the data regarding seeking professional medical advice, with ChatGPT4 showing the lowest percentage of decisive answers given, but the highest percentage of suggesting asking a doctor, and Bard showing the inverse of this. This further highlights Bard's "cocky' and almost arrogant tone in responses, and the possible increase in hallucination rates. This is also made clear by having the lowest percentage of abstaining from answering, Figure 4, implying that Bard most of the time would rather give an incorrect answer when unsure, as opposed to not giving one at all which would be safer. Further adaptations that were made upon ChatGPT3.5 in the release of ChatGPT4 were developments in contextual understanding and ambiguity resolution. To summarize this, prior to the updated model, during long conversations with lots of information input, ChatGPT3.5 would sometimes lose "focus" of the point of the question resulting in responses that were irrelevant to the original question. However, ChatGPT4 is better at this contextual understanding and allows lengthy conversations to be had that are more consistent and relevant to the actual questions asked. Furthermore, previously, ChatGPT3.5 upon not fully understanding the question, would just provide the best suited answer to the best of its understanding, which could be quite wrong. However, updates have allowed ChatGPT4 to clarify lack of understanding with follow up questions to resolve any ambiguity to ensure that the answer given is in response to a better interpreted question, allowing more realistic flow of conversation. This is relevant to this study as the high rates of decisive answers given by ChatGPT3.5, but contrasting low rates given by ChatGPT4, could suggest lack of proper understanding of these controversial scenarios and therefore production of incorrect or irrelevant responses based on a "rough" understanding. Whereas ChatGPT4's increased likelihood to attempt to clear up any confusion from the question could better reflect the controversial nature of these questions through low definitive answer rate. This is because ultimately with many of the scenarios, the management plan chosen is surgeon-dependent and how familiar and competent they are with certain techniques, depending on where and how they were trained. This is clearly understood by ChatGPT4 with a typical

response to a question following the format: brief introduction, advantages, and disadvantages of each technique, concluding statement suggesting that the decision made is surgeon dependent and therefore emphasises the importance of needing to seek real professional advice from a doctor. ChatGPT4 clearly possesses a better understanding of larger inputs of information, allowing it to recognise that the questions being asked are not simple with one clear solution, also resulting in higher numbers of abstaining from answering at all, Figure 4. However, ChatGPT3.5 in combination with this lack of developed understanding and ambiguity resolution, seems to just hallucinate instead to give a definitive answer in a lot of cases.

AI	Number of agreed responses	Agreement %	Number of abstained answers	Abstained %
ChatGPT 3.5	220	80	43	16
ChatGPT 4	243	88	79	29
Bard	213	78	30	11

Figure 4: Number of agreed responses and number of times LLM abstained from answering the question.

Once more, in line with the previous data, as expected ChatGPT4 is the most consistent with its responses and Bard is the least consistent. This could point to ChatGPT4 being more reliable as it produces the most reproducible data, with the opposite being true for Bard. Considering that ChatGPT is only based on information up to date until 2021, a further development upon ChatGPT4 with an expansion of this dataset suggests great promise for future use based on this data.

Lastly, Figure 5 (see appendix) shows a comparison of the most chosen option to each question by both the 3 LLMs combined and the surgeons. For 10 questions a consensus option could not be reached due to 3 different consensus options being generated by each AI, therefore not allowing a single consensus option to be deduced across the 3 LLMs. However, with these 10 questions excluded from the 55 questions, overall, the surgeons and LLMs agreed 23 times giving an agreement percentage of 51%. Without excluding these 10 questions, the agreement percentage falls to 42%. Since the surgeons have 153 years' worth of experience between them with constant learning, and the LLMs despite being trained on a larger dataset are only up to date until 2021 (excluding Bard which is currently up to date), this is a poor agreement rate. In my opinion, this is not surprising, and believe the more trustworthy and accurate responses to be coming from the surgeons.

CONCLUSIONS

In conclusion, whilst these LLMs have proven that they are competent in answering simple medical questions as proven by similar research papers[15], they appear to struggle with answering complex, controversial questions. This is not only in terms of accuracy, with the LLMs not always suggesting the most suitable management plan, but also in terms of providing a definitive answer at all, with a definitive and accurate response being rare. The LLMs also displayed issues with reliability since their responses for the same question were not always reproducible for each regeneration. Therefore, whilst the recent advancements in these LLMs have been impressive, allowing them to show great promise for future use within medicine, for now, it seems they are safer to be used as an educational tool, but not to be relied on solely. They can display

impressive arguments for and against certain treatment options, however, when it comes down to making a definitive decision, they cannot be trusted to provide one, let alone a safe and correct answer. Worse, they can even confidently give the incorrect solution, without the suggestion of seeking professional medical care as Bard did on multiple occasions, which could be harmful information in the hands of a patient.

APPENDIX

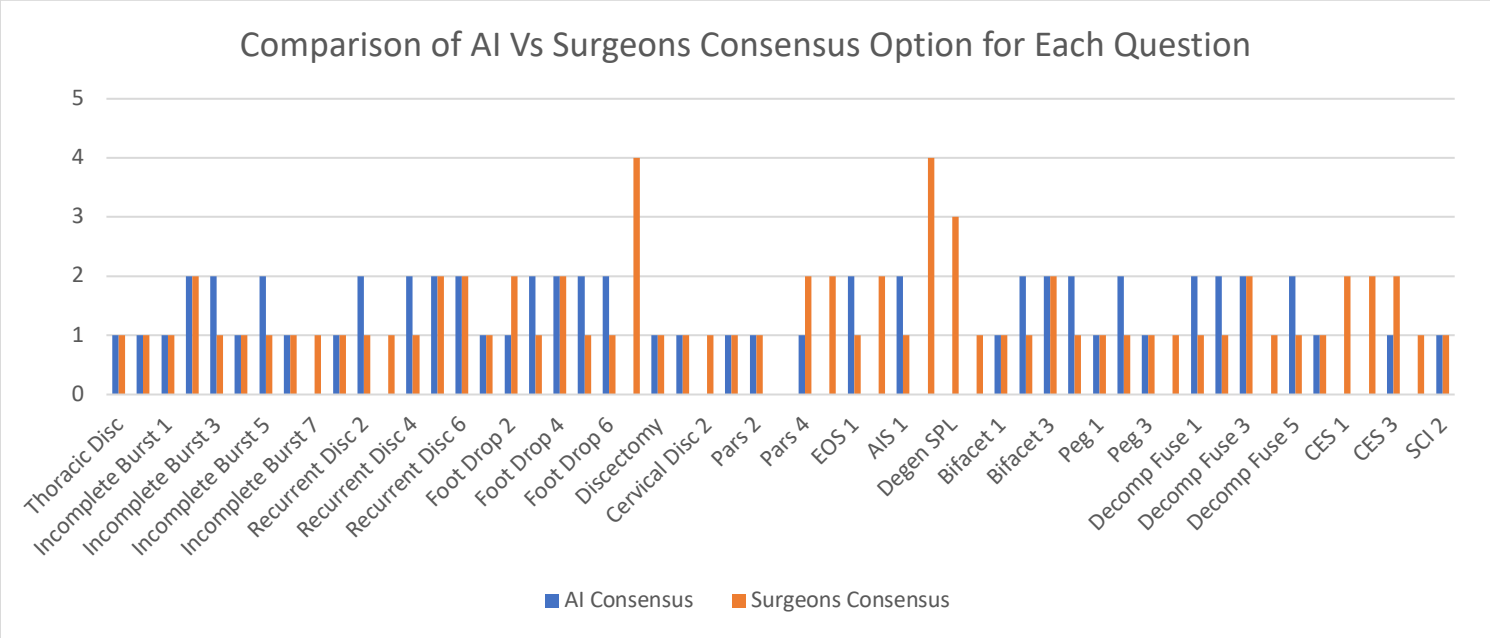

Figure 5: Comparison of AI Vs Surgeons most chosen option (consensus) for each question.

REFERENCES

1. Yao J, Burns JE, Munoz J, et al. Artificial intelligence in spinal surgery: current applications and future directions. Global Spine J. 2020;10(2):235-241.

2. Veeravagu A, Jiang B, Carvalho C, et al. Artificial intelligence and spine surgery: a narrative review. Neurospine. 2020;17(1):8-18.

3. Alimi M, Pota V, Neri E, et al. Artificial intelligence and machine learning in spine surgery: a narrative review. Neurospine. 2021;18(1):1-13.

4. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402-2410.

5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-118.

6. Nundy S, Montgomery E. Artificial intelligence and bias: four key challenges. J Gen Intern Med. 2020;35(11):3307-3309.

7. Ross C, Swetlitz I. How racially biased AI can lead to unfair treatment. Stat News. June 24, 2019. Available from: https://www.statnews.com/2019/06/24/racially-biased-artificial-intelligence-can-lead-to-unfair-treatment/. Accessed March 31, 2023.

8. Cahan P, Treutlein B. A conversation with ChatGPT on the role of computational systems biology in stem cell research. Stem Cell Reports. 2023;18(1):1-2.

9. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLoS digital health. 2023;2(2):e0000198.

10. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. arXiv preprint arXiv:221214882. 2022

11. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? The Lancet Infectious Diseases. 2023;23(4):405-6.

12. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems. 2022;35:27730-44.

13. Small J. Google exec warns of AI chatbot 'hallucinations.' Entrepreneur. Entrepreneur; 2023. [accessed 9 May 2023] Available from: https://www.entrepreneur.com/business-news/google-exec-warns-of-ai-chatbot-hallucinations/444842

14. Woodie A. Hallucinations, plagiarism, and chatgpt. Datanami. Datanami; 2023. [accessed 9 May 2023] Available from: https://www.datanami.com/2023/01/17/hallucinations-plagiarism-and-chatgpt/

15. Yeo YH, Samaan JS, Ng WH, Ting P-S, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. medRxiv. 2023:2023.02. 06.23285449.