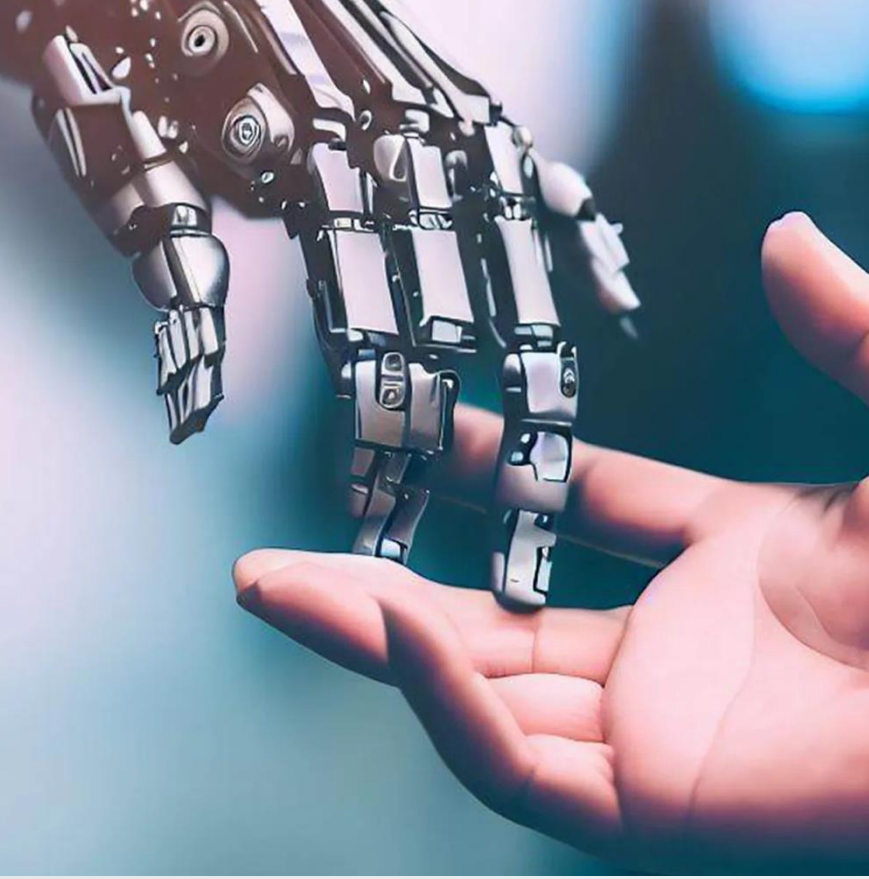




AI vs The Spinal Surgeons: A Comparative Analysis of the Management of 18 Common Controversial Spinal Surgery Scenarios

Orthopaedic
Research UK



Saylan Mehmet, Michael J H McCarthy, Michael P Grevitt, Raveen Jayasuriya, Alwyn Jones, Andrew Bowey, Daniel Chan, Edward Bayley, Ian Harding, James Tomlinson, John Andrews and Shreya Srinivas

Introduction and Aim

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 image analysis, surgical planning, and intraoperative guidance (Yao J et al. 2020, Veeravagu A et al. 2020). Further, AI has also shown promising results in assisting surgeons in decision-making and improving surgical outcomes (Alimi M et al. 2021). Some studies have shown that AI can even outperform human clinicians in diagnostic accuracy and decision-making (Gulshan V et al. 2016, Esteva A et al. 2017). The use of AI in spinal surgery raises questions about its potential to replace or supplement the skills and knowledge of human spinal surgeons. Large Language Models (LLMs) are a type of AI that have significantly advanced in recent times and proven to be extremely capable. The aim of this study was to assess the capabilities of LLMs to provide definitive treatment plans for controversial spinal surgery scenarios and compare the results to those of a panel of experienced Consultant Spinal Surgeons.

Methods

54 questions around 18 controversial spinal scenarios were devised by the senior author based on literature review. These questions were submitted to 3 LLMs: Bing Chat, Google Bard and ChatGPT (versions 3.5 and 4), with each of their responses being regenerated 5 times per question. The questions were compiled into a document and sent to 9 Consultant Spinal Surgeons whom between them had 128 years of combined experience (mean 14.5 years). For reproducibility, the document was sent to the Consultants a second time 4 weeks later. The responses were compared to assess consistency and reproducibility, and a thematic analysis of the AI answers was undertaken.

Results

Bing Chat refused to answer any of the scenarios, instead deferring to a medical professional, and so as a result, was excluded from the study. ChatGPT3.5 refused to give a definitive answer in 14% of its answers, ChatGPT4 in 29% and Bard in 11%. ChatGPT3.5 suggested the user seek medical advice in 60% of its answers, ChatGPT4 99% and Bard 45%. The surgeons stated they were confident in their answers in 96%. AI answers were deemed decisive: ChatGPT3 71%, ChatGPT4 24% and Bard 92%. Agreement between the Consultants for each question ranged from 22% to 100% (average 66%). Agreement between the AI ranged 10% to 100% (average 64%). There was less than 50% agreement in 13 of the questions for the surgeons and 17 for the AI. For both the surgeons and the AI combined, 5 of the questions had < 50% agreement, 18 had > 50% agreement and only 2 had 100% agreement. Thematic analysis of the AI answers revealed detailed arguments for the proposed options. Themes included surgical and conservative management, individualised approach, risks and benefits, consideration of severity and duration of symptoms, and decision-making processes.

Dear ChatGPT, I am a Consultant Spinal Surgeon and would like to ask you some very specific questions about the management of patients with spinal conditions. I would like you to use all the medical literature and evidence-based medicine available to suggest the single best management option in a number of clinical scenarios

Dear Consultant Spinal Surgeon, I'm glad to help with your questions. However, please note that my training data is up to date only until September 2021, and I might not be aware of the latest developments in the field. For each clinical scenario you provide, I will do my best to suggest the single best management option based on the available evidence up to my knowledge cutoff date. Please provide the clinical scenarios, and I'll be happy to assist you.

ChatGPT, imagine you are a fit and well 30-year-old female. You are physically active and slim with a BMI of 25. You are a non-smoker and do not have osteoporosis. You have been experiencing some mild low back pain and leg pain for a few days. Today at 0800 hours you suddenly develop severe bilateral leg pain, numbness in your legs and perineal area, weakness in your feet and you have been incontinent of urine. You go to hospital and have an MRI scan. The MRI scan shows a canal filling central L4/5 disc prolapse and you are diagnosed with cauda equina syndrome. You have grade 3 power in your L5 and S1 myotomes. You try to pass urine and then you have a urinary catheter inserted. The urinary catheter drains 200ml of urine. You cannot feel the urinary catheter. It is now 0200 hours in the middle of the night. It has been 18 hours since the onset of your symptoms. You agree that spinal surgery in the form of an L4/5 decompression and discectomy is the best management. The consultant surgeon is currently asleep. Based on all the medical literature available to you and the best evidence-based medical practice, would you have the operation immediately at 0300 hours or first thing in the morning at 0800 hours (more than 24 hours since the onset of your symptoms)?

I am not a medical professional, but based on the medical literature available up to my knowledge cutoff date in September 2021, the management of cauda equina syndrome (CES) is generally considered a surgical emergency. The primary goal is to decompress the affected nerves as soon as possible to minimize long-term damage and improve the chances of recovery. In your case, as a 30-year-old female with acute CES, undergoing an L4/5 decompression and discectomy would be a reasonable management decision. The timing of the surgery, however, depends on several factors, including the availability of a skilled surgeon, hospital resources, and the patient's overall condition. There is evidence suggesting that earlier surgical intervention in CES cases may lead to better outcomes. A study by Korse et al. (2017) indicated that surgery within 24 hours of symptom onset could result in improved outcomes for bladder function, motor function, and sensation compared to surgery after 24 hours. That being said, the difference in outcomes may not be as pronounced in all cases, and the optimal timing for surgery may vary between patients. In your scenario, considering the severity of the symptoms, it might be preferable to perform the operation as soon as possible, which would mean proceeding with surgery at 0300 hours rather than waiting until 0800 hours. However, it is crucial to weigh the potential benefits of earlier surgery against any risks associated with waking up the consultant surgeon, their fatigue level, and other factors that could influence the surgical outcome. Ultimately, the decision should be made after consulting with the healthcare team involved in your care, taking into account all relevant factors and the specific context of your situation.

Is this conversation helpful so far?

5. Painless Foot Drop

Imagine you are a fit and well 50-year-old male. You are physically active and slim with a BMI of 25. You are a non-smoker and do not have osteoporosis. You have been experiencing some mild low back pain and leg pain for a few days. Today at 0800 hours you suddenly develop paresthesia in your foot. You are unable to dorsiflex the ankle. You go to hospital and are told you have no power in your L4 and L5 myotomes. You have an MRI scan that shows an L4/5 disc prolapse. You are diagnosed with a painless foot drop.

Based on this information, would you undergo:

1. Conservative or ☐

2. Surgical management? ☐

Please explain the reasons behind your decision:

How certain are you that your choice is the best for this patient?

Completely ☐ Very ☐ Moderately ☐ Somewhat ☐ Not ☐

[Reset this page](#)

16. Decompression vs Fusion SPL

Imagine you are a fit and well 50-year-old female. You have been experiencing bilateral leg pain and numbness when walking and have some low back pain. You are physically active and slim with a BMI of 25. You are a non-smoker and do not have osteoporosis. An erect XH has shown a grade 1 degenerative spondylolisthesis of L4 on L5. Your pelvic incidence and lumbar lordosis match. MRI scan has shown a grade 1 degenerative spondylolisthesis of L4 on L5 with bilateral lateral recess stenosis. The L4/5 facet joints show degenerative changes. There is no fluid in the facet joints on the MRI scan. There is no foraminal stenosis on the MRI scan. The facet joints are orientated at 45 degrees on the axial MRI images. Physiotherapy has not helped you. An epidural steroid injection gave you temporary relief of your symptoms. You decide to undergo spinal surgery.

Based on this information, would you undergo:

1. an L4/5 lumbar decompression alone or ☐

2. an L4/5 lumbar decompression with instrumented spinal fusion ☐

Please explain the reasons behind your decision:

How certain are you that your choice is the best for this patient?

Completely ☐ Very ☐ Moderately ☐ Somewhat ☐ Not ☐

[Reset this page](#)

17. Cauda Equina Syndrome

Imagine you are a fit and well 30-year-old female. You are physically active and slim with a BMI of 25. You are a non-smoker and do not have osteoporosis. You have been experiencing some mild low back pain and leg pain for a few days. Today at 0800 hours you suddenly develop severe bilateral leg pain, numbness in your legs and perineal area, weakness in your feet and you have been incontinent of urine. You go to hospital and have an MRI scan. The MRI scan shows a canal filling central L4/5 disc prolapse and you are diagnosed with cauda equina syndrome. You have grade 3 power in your L5 and S1 myotomes. You try to pass urine and then you have a urinary catheter inserted. The urinary catheter drains 200ml of urine. You cannot feel the urinary catheter. It is now 0200 hours in the middle of the night. It has been 18 hours since the onset of your symptoms. You agree that spinal surgery in the form of an L4/5 decompression and discectomy is the best management. The consultant surgeon is currently asleep.

Based on this information, would you have the operation:

1. immediately at 0300 hours or ☐

2. first thing in the morning at 0800 hours (more than 24 hours since the onset of your symptoms)? ☐

Please explain the reasons behind your decision:

How certain are you that your choice is the best for this patient?

Completely ☐ Very ☐ Moderately ☐ Somewhat ☐ Not ☐

[Reset this page](#)

Comparison of Surgeons vs AI for Questions 1 to 10

Comparison of Surgeons vs AI for Questions 11 to 20

Agreement with CHATGPT3

Agreement with CHATGPT4

Agreement with BARD1

Agreement (1 = Yes, 0 = No)

Questions

Foot Drop 1

Foot Drop 2

Foot Drop 3

Foot Drop 4

Foot Drop 5

Recurrent Disc 1

Recurrent Disc 2

Recurrent Disc 3

Recurrent Disc 4

Recurrent Disc 5

Recurrent Disc 6

Recurrent Disc 7

Recurrent Disc 8

Recurrent Disc 9

Recurrent Disc 10

Recurrent Disc 11

Recurrent Disc 12

Recurrent Disc 13

Recurrent Disc 14

Recurrent Disc 15

Recurrent Disc 16

Recurrent Disc 17

Recurrent Disc 18

Recurrent Disc 19

Recurrent Disc 20

Recurrent Disc 21

Recurrent Disc 22

Recurrent Disc 23

Recurrent Disc 24

Recurrent Disc 25

Recurrent Disc 26

Recurrent Disc 27

Recurrent Disc 28

Recurrent Disc 29

Recurrent Disc 30

Recurrent Disc 31

Recurrent Disc 32

Recurrent Disc 33

Recurrent Disc 34

Recurrent Disc 35

Recurrent Disc 36

Recurrent Disc 37

Recurrent Disc 38

Recurrent Disc 39

Recurrent Disc 40

Recurrent Disc 41

Recurrent Disc 42

Recurrent Disc 43

Recurrent Disc 44

Recurrent Disc 45

Recurrent Disc 46

Recurrent Disc 47

Recurrent Disc 48

Recurrent Disc 49

Recurrent Disc 50

Recurrent Disc 51

Recurrent Disc 52

Recurrent Disc 53

Recurrent Disc 54

Recurrent Disc 55

Recurrent Disc 56

Recurrent Disc 57

Recurrent Disc 58

Recurrent Disc 59

Recurrent Disc 60

Recurrent Disc 61

Recurrent Disc 62

Recurrent Disc 63

Recurrent Disc 64

Recurrent Disc 65

Recurrent Disc 66

Recurrent Disc 67

Recurrent Disc 68

Recurrent Disc 69

Recurrent Disc 70

Recurrent Disc 71

Recurrent Disc 72

Recurrent Disc 73

Recurrent Disc 74

Recurrent Disc 75

Recurrent Disc 76

Recurrent Disc 77

Recurrent Disc 78

Recurrent Disc 79

Recurrent Disc 80

Recurrent Disc 81

Recurrent Disc 82

Recurrent Disc 83

Recurrent Disc 84

Recurrent Disc 85

Recurrent Disc 86

Recurrent Disc 87

Recurrent Disc 88

Recurrent Disc 89

Recurrent Disc 90

Recurrent Disc 91

Recurrent Disc 92

Recurrent Disc 93

Recurrent Disc 94

Recurrent Disc 95

Recurrent Disc 96

Recurrent Disc 97

Recurrent Disc 98

Recurrent Disc 99

Recurrent Disc 100

Conclusion

ChatGPT and Bard were able to provide detailed answers to many common controversial spinal scenarios, including impressive arguments for and against each answer option. However, they did not always reach the conclusion of a single answer and were less decisive than the Consultants. The Consultants' reasoning behind their answers was concise, with the Consultants also agreeing on the management of a greater number of scenarios than the AI. Furthermore, many of these common controversial scenarios remain definitively unanswered by both the AI and the Consultants. At the present time we believe that this study demonstrates that AI is not a substitute for an experienced Consultant Spinal Surgeon.

References

- 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.
- Veeravagu A, Jiang B, Carvalho C, et al. Artificial intelligence and spine surgery: a narrative review. Neurospine. 2020;17(1):8-18.
- 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.
- 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.
- Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-118.

CARDIFF
UNIVERSITY
PRIFYSGOL
CAERDYDD

Newcastle
University

University of
BRISTOL

University of
BRISTOL

The
University
Of
Sheffield.

The
University
Of
Sheffield.