ORIGINAL ARTICLE
DEFORMITIES
http://dx.doi.org/10.1590/S1808-185120252402293157

ARTIFICIAL INTELLIGENCE IN SCOLIOSIS DIAGNOSIS: A COMPARATIVE STUDY BETWEEN CHATGPT AND SURGEONS

INTELIGÊNCIA ARTIFICIAL NO DIAGNÓSTICO DA ESCOLIOSE: UM ESTUDO COMPARATIVO ENTRE CHATGPT E CIRURGIÕES

INTELIGENCIA ARTIFICIAL EN EL DIAGNÓSTICO DE LA ESCOLIOSIS: UN ESTUDIO COMPARATIVO ENTRE CHATGPT Y CIRUJANOS

Lucas Silveira Rabello de Oliveira[1], Rafael Carboni de Souza[1], André Evaristo Marcondes Cesar[1], Bruno Vieira Motter[1], Wilker Herkson de Almeida Oliveira[1], Guilherme Foizer[1], Gabrielle do Amaral Virginio Pereira[2], Luciano Miller Reis Rodrigues[1]

1. Faculdade de Medicina do ABC, Department of Orthopedics and Traumatology, Spine Surgery, São Paulo, SP, Brazil.
2. Universidade de São Paulo (USP), School of Medicine, São Paulo, SP, Brazil.

ABSTRACT

Objective: This study explores the accuracy of ChatGPT in classifying and suggesting approaches for adolescent idiopathic scoliosis, assessing the level of agreement between the artificial intelligence model's responses and the evaluations of spine surgery specialists. It aims to help answer the following question: Is it possible to trust ChatGPT-4 (natural language artificial intelligence) to recommend approaches for typical everyday cases, aiding less experienced orthopedists or even general practitioners? The proposed analysis seeks to identify the potential and limitations of applying artificial intelligence to support diagnosis and clinical decision-making without prior training of the platform. Methods: This is a cross-sectional study involving five fictitious cases of idiopathic scoliosis presented to ChatGPT, which provided the Lenke classification and a suggested approach for each case. A panel of 37 surgeons evaluated the responses, determined the best approach, and scored ChatGPT's recommendations on a Likert scale from 1 to 5, reflecting their level of agreement. Results: In simpler cases (Case 1), ChatGPT showed high agreement with the specialists, with 97.3% of the surgeons agreeing with the recommendation of "instrumentation surgery" (AC1=0.95). However, agreement was significantly lower in more complex cases (Cases 3 and 5), with only 11.1% and 18.8% of the specialists accepting the AI's recommendations, respectively. The model's accuracy in the Lenke classification was consistent across all cases, demonstrating its ability to apply standardized criteria. There was no significant correlation between the surgeons' experience and their level of agreement with the software. Conclusion: ChatGPT showed potential as an auxiliary tool in the diagnosis and therapeutic planning of scoliosis, particularly in classification, but it is not yet ready to be used reliably and consistently, especially in more complex cases, particularly when considering clinical nuances and individual patient factors. Although promising, the adoption of this technology can complement clinical judgment but still requires supervision and does not replace the role of specialized medical evaluation in the current scenario. ***Level of Evidence IV; Descriptive Observational Studies.***

Keywords: Artificial Intelligence; Scoliosis; Spine; Spinal Fusion; Spinal Curvatures; Comparative Study.

RESUMO

Objetivo: Este estudo explora a acurácia do ChatGPT na classificação e sugestão de condutas para escoliose idiopática do adolescente, avaliando o nível de concordância entre as respostas do modelo de inteligência artificial e as avaliações de especialistas em cirurgia de coluna vertebral. Procurando ajudar a responder a seguinte pergunta: É possível confiar no ChatGPT-4 (inteligência artificial de linguagem natural) para recomendar condutas diante de casos típicos do cotidiano, servindo como um auxílio para ortopedistas menos experientes ou até mesmo para médicos generalistas? A análise proposta busca identificar o potencial e as limitações da aplicabilidade da inteligência artificial no suporte ao diagnóstico e decisão clínica, sem treinamento prévio da plataforma. Métodos: Trata-se de um estudo transversal com cinco casos fictícios de escoliose idiopática apresentados ao ChatGPT, que forneceu a classificação de Lenke e uma sugestão de conduta para cada caso. Um painel de 37 cirurgiões avaliou as respostas, emitiu a melhor conduta e pontuou as recomendações do ChatGPT em uma escala Likert de 1 a 5, refletindo seu nível de concordância. Resultados: Em casos mais simples (Caso 1), o ChatGPT demonstrou alta concordância com os especialistas, com 97,3% dos cirurgiões concordando com a recomendação de "cirurgia de instrumentação" (AC1=0,95). Entretanto, em casos mais complexos (Casos 3 e 5), a concordância foi significativamente menor, com apenas 11,1% e 18,8% dos especialistas aceitando as recomendações da IA, respectivamente. A precisão do modelo na classificação de Lenke foi consistente em todos os casos, evidenciando sua capacidade de aplicar critérios padronizados. Não houve correlação significativa entre a experiência dos cirurgiões e o nível de concordância com o software. Conclusão: O ChatGPT demonstrou potencial como ferramenta auxiliar no diagnóstico e planejamento terapêutico de escoliose, especialmente na classificação, porém ainda não está pronta para ser utilizada de maneira confiável e replicável, especialmente em casos mais complexos, especialmente em considerar nuances clínicas e fatores individuais do paciente. Apesar de promissora, a adoção dessa tecnologia poderá complementar o julgamento clínico, mas ainda requer supervisão e não substitui o papel da avaliação médica especializada no cenário atual. ***Nível de Evidência IV; Estudos Observacionais Descritivos.***

Descritores: *Inteligência Artificial; Escoliose; Coluna Vertebral; Fusão vertebral; Curvaturas da Coluna Vertebral; Estudo Comparativo.*

Study conducted by the Hospital Estadual Mário Covas, Dr. Henrique Calderazzo Street, 321, Santo André, SP, Brazil. 09190-615.
Correspondence: Lucas Silveira Rabello de Oliveira. 321, Dr. Henrique Calderazzo Street, Santo André, SP, Brazil. 09190-615. doutorrabello@gmail.com

CC BY

Received on 30/01/2025 accepted on 12/03/2025
Reviewed by: Marcelo Wajchenberg

RESUMEN

Objetivo: Este estudio explora la precisión de ChatGPT en la clasificación y sugerencia de conductas para la escoliosis idiopática del adolescente, evaluando el nivel de concordancia entre las respuestas del modelo de inteligencia artificial y las evaluaciones de especialistas en cirugía de columna vertebral. Busca ayudar a responder la siguiente pregunta: ¿Es posible confiar en ChatGPT-4 (inteligencia artificial de lenguaje natural) para recomendar conductas en casos típicos del día a día, sirviendo como una ayuda para ortopedistas menos experimentados o incluso médicos generalistas? El análisis propuesto busca identificar el potencial y las limitaciones de la aplicabilidad de la inteligencia artificial en el apoyo al diagnóstico y la toma de decisiones clínicas, sin entrenamiento previo de la plataforma. Métodos: Se trata de un estudio transversal con cinco casos ficticios de escoliosis idiopática presentados a ChatGPT, que proporcionó la clasificación de Lenke y una sugerencia de conducta para cada caso. Un panel de 37 cirujanos evaluó las respuestas, emitió la mejor conducta y puntuó las recomendaciones de ChatGPT en una escala Likert de 1 a 5, reflejando su nivel de concordancia. Resultados: En casos más simples (Caso 1), ChatGPT demostró una alta concordancia con los especialistas, con el 97,3% de los cirujanos de acuerdo con la recomendación de "cirugía de instrumentación" (AC1=0,95). Sin embargo, en casos más complejos (Casos 3 y 5), la concordancia fue significativamente menor, con solo el 11,1% y el 18,8% de los especialistas aceptando las recomendaciones de la IA, respectivamente. La precisión del modelo en la clasificación de Lenke fue consistente en todos los casos, evidenciando su capacidad para aplicar criterios estandarizados. No hubo una correlación significativa entre la experiencia de los cirujanos y el nivel de concordancia con el software. Conclusión: ChatGPT demostró potencial como herramienta auxiliar en el diagnóstico y la planificación terapéutica de la escoliosis, especialmente en la clasificación, pero aún no está listo para ser utilizado de manera confiable y replicable, particularmente en casos más complejos, especialmente al considerar matices clínicos y factores individuales del paciente. Aunque prometedora, la adopción de esta tecnología puede complementar el juicio clínico, pero aún requiere supervisión y no reemplaza el papel de la evaluación médica especializada en el escenario actual. ***Nivel de Evidencia IV; Estudios Observacionales Descriptivos.***

Descriptores: *Inteligencia Artificial; Escoliosis; Columna Vertebral; Fusión Vertebral; Curvaturas de la Columna Vertebral; Estudio Comparativo.*

INTRODUCTION

Artificial Intelligence (AI) has attracted increasing attention in the medical field, promising radical transformations in healthcare delivery. However, the successful implementation of AI in clinical practice requires a rigorous evaluation of its effectiveness and reliability.[1]

The term "machine learning" was coined by Arthur Samuel in 1959,[2] defining a field of artificial intelligence (AI) in which computers learn automatically from the accumulation of data. This technique has been widely applied to analyzing large volumes of data, particularly for image processing.[3] Unlike traditional software, which requires specific instructions to perform a task, deep learning allows systems to recognize patterns autonomously and make predictions.[4]

According to Pesapane et al, the application of AI will transform working methodologies in various professions, including medicine, and in radiology this change will occur more rapidly than in other medical specialties.[5]

ChatGPT is a very advanced natural language artificial intelligence, to become a large language model (LLM), developed by OpenAI. It is the most widely used today, and perhaps, in its current version (ChatGPT-4o), it is the most complete and accurate in this proposal model.[6]

The model was learned from various texts, such as books, articles, websites, and other educational and scientific materials. However, it is important to mention that although the model can generate informed answers based on the information it has learned, it does not have access to individual, confidential, or up-to-date medical data. Generative AIs generate responses based on the probabilistic relationships between the words and phrases in their training base.[6,7]

ChatGPT does not have knowledge or awareness on its own but uses learned patterns to produce relevant and coherent answers; It should also be noted that the software has limitations in terms of critical analysis of the information presented.[1,6] Therefore, medical conduct is extremely individualized, depending on various factors intrinsic to the patient and the experience acquired in the career of each surgeon. We consider it unlikely that ChatGPT will provide reliable information that allows one to postpone or even dispense with a specialized medical evaluation.

With the advent of artificial intelligence and its growing application in various fields, the impact of AI technology, particularly the ChatGPT language model, on medical and surgical practice has become a crucial topic of investigation. Specifically, the application of this advanced technology in spinal surgery is the focus of this study. In addition, He et al. noted that ChatGPT can improve intraoperative support by providing real-time surgical navigation information and monitoring of physiological parameters, as well as assisting in postoperative rehabilitation guidance, optimizing the collection and analysis of patient data, improving communication between spine surgeons, patients and their families, and contributing to the surgical planning process.[8]

The study explores the accuracy of classifying and suggesting procedures for adolescent idiopathic scoliosis, evaluating the level of agreement between the chatbot's responses and the assessments of spinal surgery specialists. Trying to help answer the following question: Is it possible to rely on ChatGPT-4 (natural language artificial intelligence) to recommend courses of action for typical everyday cases, aiding less experienced orthopedists or even general practitioners? The proposed analysis seeks to identify the potential and limitations of using artificial intelligence in clinical practice, contributing to the existing literature. This allows us to evaluate the applicability of artificial intelligence in supporting diagnosis and clinical decision-making in the current scenario without prior training of the platform.

METHOD

The study uses a cross-sectional investigation using a quantitative methodology to evaluate the agreement between ChatGPT-4o and spinal surgery specialists in five fictitious Adolescent Idiopathic Scoliosis (AIS) cases. The analysis focused on comparing the treatment options suggested by professionals and by Artificial Intelligence (AI), as well as the level of agreement of surgeons with the explanations provided by AI. The AI analyzed the cases presented and provided a Lenke classification and a suggested course of action for each case. A panel of experts was asked to rate each ChatGPT-4 answer using a 5-point Likert scale, ranging from "strongly disagree" to "strongly agree".

Data collection

Data was collected using an online questionnaire containing the clinical cases to be analyzed and the ChatGPT4 responses regarding pathology classifications and the suggested medical approach. The answers were analyzed anonymously and published, with the data grouped without identifying the participants. Furthermore, as this is a study without clinical intervention and observational methodology, based on the security of the information and without identifying the patients, there will be no additional risk promoted by this research. All the researchers involved in this work, individually and collectively, undertook to use the data from this research only for descriptive purposes, respecting the secrecy and confidentiality of the data collected.

Thirty-seven surgeons, including orthopedic spine surgery specialists, spine surgery fellows, neurosurgery residents, and neurosurgeons recruited from various healthcare institutions, took part in the study. Each participant received five detailed clinical cases of AIS, including information necessary for Lenke's classification and the therapeutic decision.

All participants signed and agreed to the ICF under CNS Resolution No. 466 of 2012. This study was submitted to the Brazil Platform and evaluated by the Research Ethics Committee: 7.070.552 and the CAAE: 81977424.5.0000.0082. Following the provisions of CNS Resolution No. 510 of April 7, 2016, the project was exempted from ethical analysis and did not require further evaluation by the committee.

After each case was presented, the surgeons were asked to choose the best treatment option from the following alternatives: a) TLSO vest; b) CTLSO vest; c) instrumentation surgery; d) expectant - with serial follow-up; e) expectant - discharge.

The participants then read the ChatGPT response, which included Lenke's classification and the treatment proposal. Finally, the surgeons assessed their level of agreement with the AI's response on a Likert scale from 1 to 5, where 1 = strongly disagree; 2 = disagree; 3 = partly agree; 4 = agree; 5 = strongly agree.

Data Analysis

Analyses were carried out using Jamovi[9,10] and Excel statistical software. Categorical variables are presented as frequency and percentages (%) and continuous variables as mean and standard deviation (M±SD). Values of $p<0.05$ are considered statistically significant.

To analyze the data, we calculated the absolute and relative frequencies of the surgeons' treatment options in each case, comparing them with the option suggested by ChatGPT. The answers "TLSO vest" and "CTLSO vest" were grouped as "vest treatment", as both represent similar orthotic treatment modalities.

Due to the nominal nature of the data and the number of evaluators, we used Gwet's AC1 Coefficient to assess the agreement between the surgeons and the AI. This coefficient is suitable for categorical data with multiple raters and is less susceptible to the prevalence paradox that can affect other coefficients, such as Fleiss's Kappa. The AC1 Coefficient ranges from 0 to 1, where values close to 1 indicate high agreement beyond chance, and values close to 0 indicate agreement equivalent to chance. Its interpretation is similar to that of the Kappa coefficient, with values between 0.81 and 1.00 indicating almost perfect agreement and between 0.61 and 0.80 indicating substantial agreement.

To assess the consistency of the surgeons' clinical decisions, we calculated the proportion of agreement observed (P_o) in each case. This metric reflects the degree of clinical consensus, indicating how often surgeons have chosen the same treatment option for a specific case. A high P_o indicated a more established clinical consensus, while a lower P_o indicated greater variability in treatment preferences. This analysis made it possible to identify agreement patterns between the experts and compare the variability according to the complexity of each clinical case.

The surgeons evaluated the ChatGPT answers on a Likert scale from 1 to 5 to indicate their level of agreement with the AI recommendations. For each case, measures of central tendency (mean, median, mode) and dispersion (standard deviation and interquartile range) were calculated to make it easier to visualize agreement patterns and identify potential biases in responses.

We investigated whether the surgeons' experience influenced the level of agreement with ChatGPT, calculating the Spearman correlation between years of experience and the level of agreement in each case. Spearman's correlation, suitable for non-parametric data, assesses the strength and direction of the association between two ordinal or interval variables.

We identified cases in which surgeons chose the same treatment option as the AI but disagreed and developed hypotheses to explain these discrepancies. Finally, we used the Kruskal-Wallis test to assess whether professional training (orthopedist, fellow, or neurosurgeon) influenced the level of agreement.

RESULTS

General data analysis

Table 1 shows the years of experience of the 37 professionals participating in the study. The group was made up of 9 professionals (24.3%) with less than two years' experience in spinal surgery, nine professionals (24.3%) with between two and five years' experience, 10 professionals (27%) with between five and ten years' experience and nine professionals (24.3%) with more than ten years' experience. The homogeneous distribution of participants showed no statistically significant difference between the proportions of professionals in the different experience groups ($\chi^2 = 0.0811$, $p = 0.994$), indicating that the sample is representative and balanced.

Case 1

In Case 1, ChatGPT recommended "instrumentation surgery" as the best treatment option. Of the 37 surgeons, 36 (97.3%) also selected "instrumentation surgery", while only one (2.7%) chose "TLSO vest". Gwet's AC1 Coefficient between the surgeons and ChatGPT was 0.95, indicating almost perfect agreement. The proportion of direct agreement was 97.3%, reflecting a very high level of agreement between the surgeons and ChatGPT. (Figure 1)

The proportion of agreement between surgeons observed (P_o) between surgeons in this case was 0.946, indicating high agreement. This means that 94.6% of the possible combinations of answers between the surgeons agreed. This high level of agreement reflects a well-established clinical consensus on the appropriate treatment for Case 1.

The average level of agreement on the Likert scale was 4.27 (standard deviation = 0.769150), with a median of 4.0 and a mode of 5. The Interquartile Range (IQR) was 1.0 (Q1 = 4.0; Q3 = 5.0), with variance: 0.59. No surgeon gave grades at levels 1 or 2, reinforcing the general acceptance of the ChatGPT recommendation. The bar graph (Figure 2) shows that most surgeons gave high marks for agreement (4 and 5), reinforcing this trend.

There was a significant negative correlation between years of experience and level of agreement (rho = -0.29; p = 0.040), suggesting that surgeons with less experience tend to agree more with ChatGPT.

CASE 2

In Case 2, ChatGPT recommended a "TLSO vest" as the best treatment option. When "TLSO vest" and "CTLSO vest" were grouped as "vest treatment", 33 of the 36 surgeons (91.7%) agreed with AI's recommendation. The other three surgeons (8.3%) chose

Table 1. Study participants (N=37).

Years of experience in spinal surgery	n	%
Less than 2 years	9	24.3%
Between 2 and 5 years	9	24.3%
Between 5 and 10 years	10	27.0%
More than 10 years	9	24.3%

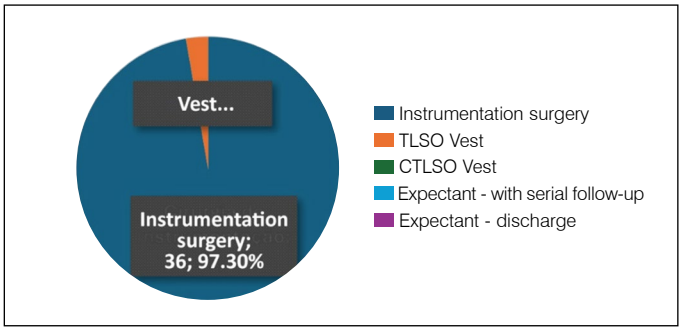

Figure 1. Proportion of participants' responses to Case 1.

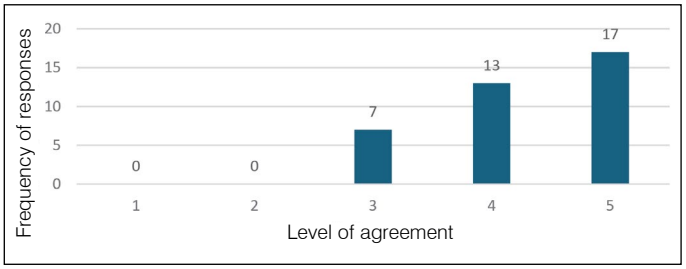

Figure 2. Proportion of Likert scale responses for case 1.

different options, such as "instrumentation surgery" or "expectant with serial follow-up". Gwet's AC1 coefficient between the surgeons and ChatGPT was 0.86, indicating almost perfect agreement. The proportion of direct agreement was 91.7%. (Figure 3)

The proportion of agreement between surgeons observed (P_o) between the surgeons in this case was 0.722, indicating substantial agreement. This means that 72.2% of the combinations of answers between the surgeons were in agreement.

The average level of agreement on the Likert scale was 2.94 (standard deviation = 1.433167), with a median of 3.5 and a mode of 4.0. The Interquartile Range (IQR) was 2.25 (Q1 = 1.75; Q3 = 4.0), with variance: 2.05, indicating that the most frequent answer was "partially agree", showing that surgeons' opinions varied, although there was a tendency to agree. (Figure 4)

There was no significant correlation between the surgeons' years of experience and the level of agreement with ChatGPT ($\rho = -0.09$; $p = 0.58$).

Case 3

In Case 3, ChatGPT suggested a "TLSO vest" as a treatment. Of the 36 surgeons who responded to this case, 32 (88.9%) chose "instrumentation surgery", while 4 (11.1%) opted for "treatment with a vest" and none chose "expectant". Gwet's AC1 coefficient between the surgeons and ChatGPT was 0.03, indicating disagreement. (Figure 5)

The proportion of direct agreement between the surgeons was 88.9%, reflecting a high level of agreement in the choice of "instrumentation surgery". The proportion of agreement observed (P_o) between the surgeons in Case 3 was approximately 0.792, indicating

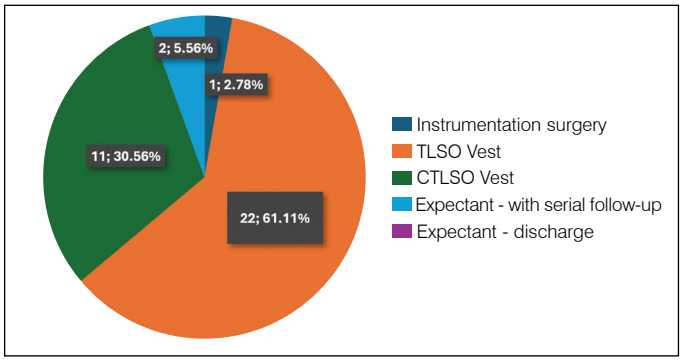

Figure 3. Proportion of participants' responses to Case 2.

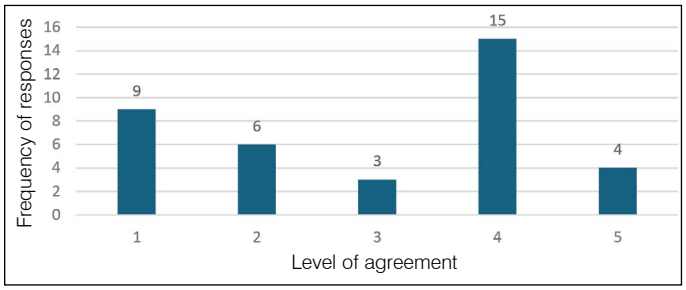

Figure 4. Proportion of Likert scale responses for case 2.

substantial to almost perfect agreement. This means that 79.2% of the possible combinations of answers between the surgeons agreed, reflecting the high level of agreement in choosing "instrumentation surgery" as the appropriate treatment.

The average level of agreement on the Likert scale was 1.89 (standard deviation = 0.863174), with a median of 2.0 and a mode of 2. The Interquartile Range (IQR) was 1.0 (Q1 = 1.0; Q3 = 2.0), with variance: 0.74, with the most frequent answers tending towards disagreement. Most surgeons gave low scores (1 or 2), reflecting general disagreement with the AI recommendation (Figure 16). There was no significant correlation between years of experience and level of agreement ($\rho = -0.05$; $p = 0.71$). (Figure 6)

Case 4

The surgeons' choices were: "instrumentation surgery" by 10 surgeons (30.3%), "TLSO vest" by 1 surgeon (3.0%), "CTLSO vest" by 1 surgeon (3.0%), "expectant - serial follow-up" by 21 surgeons (63.6%) and "expectant - discharge" by 4 surgeons (10.8%). Therefore, 10 surgeons (30.3%) agreed with the AI recommendation. (Figure 7)

The proportion of agreement observed (P_o) between the surgeons was 0.393, indicating moderate agreement. This means that 39.3% of the combinations of answers between the surgeons were in agreement.

The average level of agreement on the Likert scale was 2.30 (standard deviation = 1.076643), with a median of 2.0 and a mode

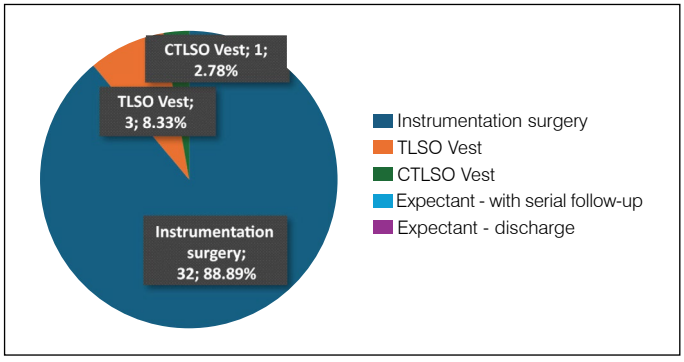

Figure 5. Proportion of participants' responses to Case 3.

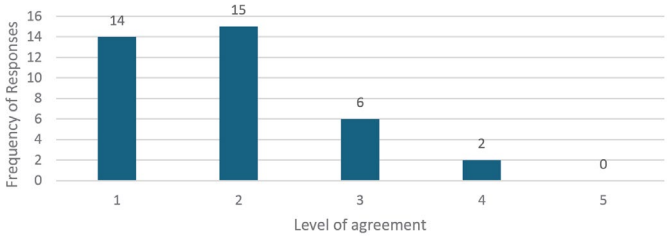

Figure 6. Proportion of Likert scale responses for case 3.

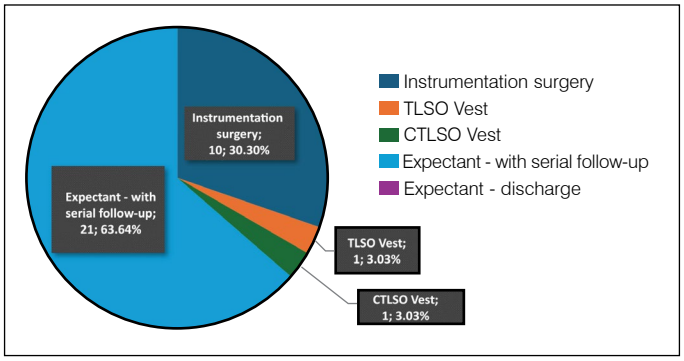

Figure 7. Proportion of participants' responses to Case 4.

of 2. The Interquartile Range (IQR) was 1.0 (Q1 = 2.0; Q3 = 3.0), with variance: 1.159, reflecting greater variability and a tendency to disagree with the AI. There was no significant correlation between experience and agreement (ρ = -0.07; p = 0.63). (Figure 8)

Case 5

Considering "TLSO vest" and "CTLSO vest" as treatment with a vest, 6 surgeons (18.8%) agreed with AI's suggestion. The majority disagreed with AI: 15 participants opted for expectant treatment - with serial follow-up, representing 46.9%, while 11 participants (34.4%) opted for expectant treatment - discharge. Gwet's AC1 coefficient was 0.18, indicating slight agreement. (Figure 9)

The proportion of agreement observed (P_o) between the surgeons in Case 5 was approximately 0.721, indicating substantial agreement. This means that 72.1% of the possible combinations of answers between the surgeons were in agreement. Although there was a majority who opted for "expectant", the presence of varied answers resulted in slightly lower agreement compared to Case 3.

The average level of agreement on the Likert scale was 1.81 (standard deviation = 1.221061), with a median of 1.0 and a mode of 1. The Interquartile Range (IQR) was 1.0 (Q1 = 1.0; Q3 = 2.0), with variance: 1.49, indicating low variability, but the tendency to disagree with the AI was even more evident. (Figure 10)

Compiled Case Analysis

Table 2 shows the response pattern of the experts in the cases presented and the degree of agreement between the experts and the ChatGPT-4 response.

Table 3 and Figure 11 show the experts' perception of whether they agree with the answer given by ChatGPT-4 via a Likert Scale.

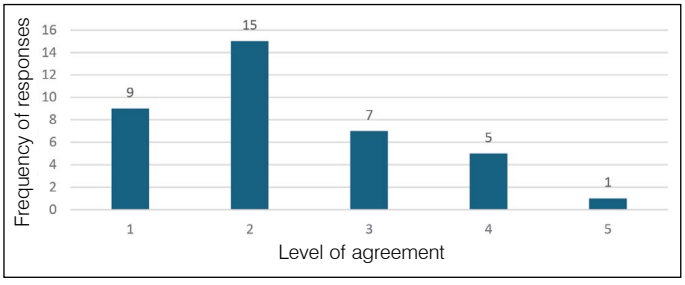

Figure 8. Proportion of Likert scale responses for case 4.

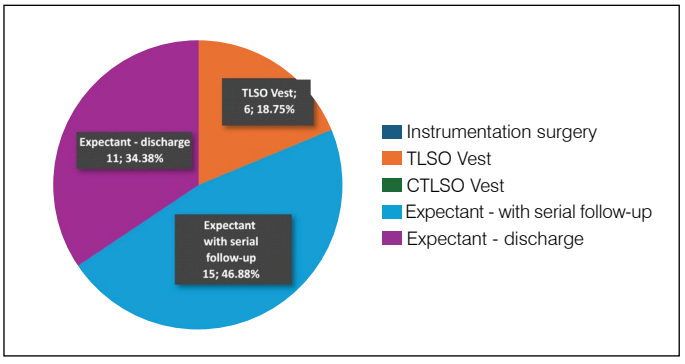

Figure 9. Proportion of participants' responses to Case 5.

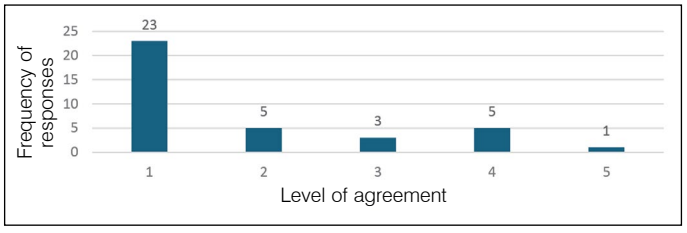

Figure 10. Proportion of Likert scale responses for case 5.

Table 2. The experts' response pattern in the cases presented and the percentage of agreement between the experts' solutions and ChatGPT-4 (N=37).

	N (%)				
	Case 1	**Case 2**	**Case 3**	**Case 4**	**Case 5**
TLSO vest	1 (2.7%)	22 (61.1%)	3 (8.3%)	1 (3.0%)	6 (18.8%)
CTLSO vest	0 (0%)	11 (30.6%)	1 (2.8%)	1 (3.0%)	0 (0%)
Surgery	36 (97.3%)	1 (2.8%)	32 (88.9%)	10 (30.3%)	0 (0%)
Exp - FU.	0 (0%)	2 (5.6%)	0 (0%)	21 (63.6%)	15 (46.9%)
Exp – D.	0 (0%)	0 (0%)	0 (0%)	0 (0%)	11 (34.4%)
Total	37 (100%)	36 (100%)	36 (100%)	33 (100%)	32 (100%)
% Agreement with ChatGPT-4	97.3% (Surgery)	91.7% (Orthosis)	11.1% (Orthosis)	30.3% (Surgery)	18.8% (Orthosis)

Table 3. Experts' perception of agreement using the Likert scale (N=37).

	Likert Scale						
	n	**IVC**	**CVC**	**Mean ± SD**	**Fashion**	**Median**	**IC (95%)**
Case 1	37	0.81	0.85	4.3±0.8	5	4	4.01-4.53
Case 2	36	0.50	0.59	2.9±1.4	4	3.5	2.46-3.43
Case 3	37	0.05	0.38	1.9±0.9	2	2	1.60-2.18
Case 4	37	0.16	0.46	2.3±1.1	2	2	1.94-2.66
Case 5	37	0.16	0.36	1.8±1.2	1	1	1.40-2.22
P*				<0.001			

* One-way ANOVA test for heterogeneous samples.

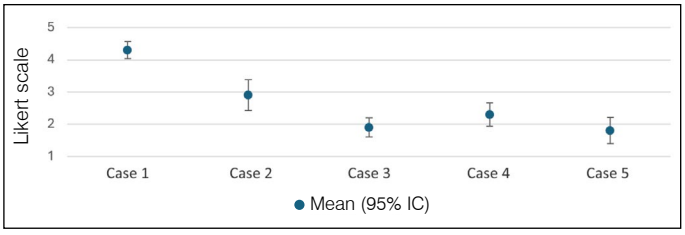

Figure 11. Experts' perception of agreement with ChatGPT-4 using a Likert scale. Test performed: One-way ANOVA (F_{Welch}=50,7, p<0,001). Games-Howell post-hoc test shows statistical differences between Case 1 and all the other cases (p<0.001) and Case 2 in relation to Cases 3 (p=0.003) and 5 (p=0.005).

We observed a statistically significant difference between the cases using the One-way ANOVA test for heterogeneous samples (F_{welch}=50.7, *p<0.*001).

The Game-Howell post-test shows that the perception of agreement for Case 1 is significantly higher than for all the other cases (*p<0.*001) and Case 2 had a higher perception of agreement than Cases 3 (*p=0.*003) and 5 (*p=0.*005). The Content Validity Index (CVI) and Content Validity Coefficient (CVC) show that only Case 1 presented an acceptable value between the experts' level of agreement with ChatGPT-4. Acceptable agreement values are those greater than 0.8 and preferably greater than 0.9. Calculating Fleiss's Kappa coefficient considering all five cases, we obtained a value of 0.552, indicating moderate agreement between the surgeons throughout the cases.

Influence of professional experience

No significant correlations were found between professional experience and the level of agreement with ChatGPT in any of the cases. Spearman's correlation coefficients ranged from -0.12 to -0.03, with p-values above 0.05. This suggests that the surgeons' experience did not have a significant impact on their agreement with the AI recommendations.

The perception of agreement was checked using a Likert scale according to the expert's years of experience. We observed a statistically significant difference only in the perception of experts in relation to Case 1 (p=*0*.004). In Case 1, experts with less than 2 years' experience had higher average values on the Likert scale compared to those with between 5 and 10 years' experience (p=0.041) and those with more than 10 years' experience(p=0.021). (Table 4 and Figure 12)

Table 4. Experts' perception of agreement using the Likert scale by years of experience (N=37).

		Likert Scale				
	n	**Case 1**	**Case 2**	**Case 3**	**Case 4**	**Case 5**
Less than 2 years	9	4.89±0.33	2.56±1.42	1.78±0.67	2.78±0.97	1.78±1.56
Between 2 and 5 years	9	4.11±0.93	3.89±0.93	2.33±1.00	2.22±1.09	2.33±1.41
Between 5 and 10 years	10	4.20±0.63	2.67±1.58	1.50±0.71	2.30±1.16	1.70±1.16
More than 10 years	9	3.89±0.78	2.67±1.50	2.00±1.00	1.89±1.05	1.44±0.53
p*		0.004	0.073	0.257	0.370	0.398

* One-way ANOVA test for heterogeneous samples

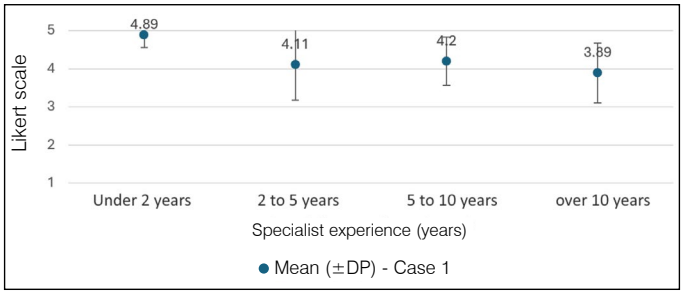

Figure 12. Experts' perception of agreement with ChatGPT-4 in Case 1 using a 5-point Likert scale by years of experience of the expert. Test performed: One-way ANOVA (F=6,39, *p*=0,004). Games-Howell post-hoc test shows statistical differences between experts with less than two years' experience and those with between five and ten years' experience (*p*=0.041) and those with more than ten years' experience(*p*=0.021).

DISCUSSION

Although it has access to limited medical data, ChatGPT, incredibly, achieved a performance equivalent to that of a third-year medical student in medical licensing exams widely applied in the United States, which has stimulated urgent discussions in the medical field.[11]

According to Stokel-Walker and van Noorden, ChatGPT could answer some medical questions openly almost as well as an average human doctor, although it still had shortcomings and inaccuracies.[12]

However, despite the vast possibilities, we also have to confront the limitations and ethical challenges that arise with this technology. With particular emphasis on the potential for threats to data security and privacy, the proper and supervised use of ChatGPT/GPT-4 is of paramount importance, leading to the conclusion that ChatGPT/GPT-4, when used carefully and responsibly, can be a useful beacon for spine surgeons.[8] One should not ignore the issues of inaccuracy of the data generated, misconduct and plagiarism, which demand attention.[13]

In telemedicine, information is generally limited to language,[13] making it potentially suitable for interventions using LLMs. However, ChatGPT is unable to ask questions to seek further clarification on questions or scenarios.

Based on the observations of Howard et. al. on questions posed to ChatGPT about clinical situations involving conduct and the decision-making process of antimicrobial choice, the chatbot was able to recognize the importance of clinically relevant factors when explicit information was provided, but missed relevant questions in scenarios of increasing complexity. The spelling and grammar in the answers were consistent, and the meaning was clear. This study concludes that the biggest obstacles to implementing this AI in clinical practice are deficiencies in situational awareness and inference. These failures can put patient safety at risk. ChatGPT seems to have access to sufficient training data, although it does not have access to specific medical databases. Even without specific training in clinical counseling, he provides convincing answers to most requests.[14]

The results of this study provide important insights into the interaction between artificial intelligence and clinical practice in the context of EIA. Case 1, which was less complex, showed the highest level of agreement between the experts and ChatGPT-4 (97%), while Cases 3 and 5 showed the lowest levels of agreement (8% and 19% respectively).

ChatGPT has demonstrated its ability to interpret clinical data and provide recommendations in line with established practices in simple cases, as evidenced by the high level of agreement in Case 1, both among survey participants and in relation to AI. The high average on the Likert scale (4.27) and the high agreement coefficients (Gwet's AC1 Coefficient between the surgeons and ChatGPT was 0.95) suggest a well-established clinical consensus and confidence in the AI recommendation for this case. The proportion of agreement observed (P_o) of 0.946 between the surgeons in Case 1 indicates a solid clinical consensus. Previous studies corroborate that AI can be effective in supporting standardized clinical decisions.[15-17]

In the more complex cases (Cases 3 to 5), agreement decreased significantly. This can be attributed to the variability in treatment options in complex clinical situations and the need for refined clinical judgment. AI, in its current form, may not be able to capture subjective and individualized factors that influence decision-making, such as patient preferences, socioeconomic context and accumulated clinical experience.[1,7,18] For example, in Case 3, although the proportion of agreement between surgeons was substantial (P_o ≈ 0.708), indicating a consensus between professionals, the surgeons' high disagreement with ChatGPT highlights the limitations of AI in complex contexts.

The general agreement between the surgeons indicates moderate agreement over the five cases. This suggests that although there is consensus in some cases, there is variability in clinical decisions, possibly influenced by the complexity of cases and individual preferences. As in the example of case 5, where there was a variability of responses in behaviors, with only 18.8% of participants agreeing with AI.

The absence of a significant correlation between surgeons' experience and the level of agreement with ChatGPT indicates that disagreement with AI is not related to the level of professional experience. Both less experienced and more experienced surgeons have been shown to critically evaluate AI recommendations, especially in complex cases that require refined clinical judgment.

Situations were identified in which surgeons chose the same treatment option as ChatGPT, but gave low levels of agreement on the Likert scale. This suggests that although the treatment option was the same, the reasoning or specific details presented by the AI were not fully accepted.

ChatGPT's accuracy in Lenke's classification demonstrates its ability to interpret objective clinical information and apply complex criteria consistently. This is particularly relevant, considering that Lenke's classification is fundamental in the surgical planning of AIS.[19,20] AI's ability to process and apply standardized classifications can help standardize conduct and reduce variability in clinical decisions,[21] especially with the advancement of machine learning technologies, considered a valuable tool for pre-surgical planning, intraoperative guidance, neurophysiological monitoring and predicting neurosurgical outcomes.[18,21]

The discrepancies identified between the choice of treatment and the level of agreement suggest that agreeing on the choice of treatment does not necessarily imply agreeing on the details or the underlying reasoning. For example, the AI may recommend different levels of arthrodesis from those that surgeons consider ideal, which affects the overall evaluation of the recommendation. Furthermore, differences in the interpretation of clinical data or the application of

evidence can lead to disagreements, even when the treatment option is the same. Furthermore, the AI may have presented inadequate concepts or technical details that do not correspond to current clinical practice. This highlights the importance of AI providing not only the recommendation, but also a clear and reasoned explanation.[22]

Some surgeons prefer TLSO even for curves above T8 due to better treatment compliance, while others opt for CTLSO in line with traditional recommendations. Studies indicate that the TLSO is effective in controlling the progression of adolescent idiopathic scoliosis (AIS) and may be more comfortable, less restrictive and less conspicuous in appearance, which may improve adherence to wearing the brace.[23,24]

Considering that the options "TLSO vest" and "CTLSO vest" are similar, especially in the clinical context, we grouped these answers as "treatment with vest". This decision reflects clinical practice, where the choice of vest type can vary according to the surgeon's experience and the patient's preferences. The variability in vest treatment choices may highlight the need to personalize AI recommendations.[20]

The findings of this study suggest that ChatGPT can be a valuable auxiliary tool for clinical decision support in cases of AIS with established consensus. However, in complex cases, the variability in surgeons' opinions highlights the importance of individual clinical judgment. AI should be seen as a complementary tool that can enrich the decision-making process, but it does not replace the expertise and reasoning of the healthcare professional.

Future research should explore the integration of AI into the clinical workflow, developing strategies so that tools such as ChatGPT complement, rather than replace, medical reasoning, making them capable of integrating more detailed clinical information, including patient satisfaction and treatment effectiveness. In addition, it is essential to work on improving AI models so that they can take into account individual patient factors and offer more personalized recommendations. Education and training of health professionals in the use of these technologies are also essential to maximize their benefits and minimize potential risks.[25]

This study has some limitations. The use of fictitious cases may not fully reflect the complexity and variability of real cases. The sample size, although adequate for exploratory analyses, may limit the generalizability of the results.

In this research journey, we break new ground for the practice of medicine and surgery, offering insights into how artificial intelligence, particularly ChatGPT/GPT-4, can become an invaluable tool. This work is an important milestone in understanding the impact of this advanced technology on spinal surgery and shaping the future of surgical practice.

CONCLUSION

This study showed that ChatGPT is capable of interpreting clinical data and providing recommendations in line with established practices in cases of AIS. High agreement in simple cases indicates potential for use as clinical decision support.

However, the variability observed in complex cases and the influence of professional experience highlight the importance of clinical judgment and individualized treatment. AI should be seen as a complementary tool that can enrich the decision-making process, but does not replace the expertise and reasoning of the healthcare professional. Especially in the current context, where there are still many errors in interpreting concepts and prioritizing clinical data.

The effective integration of AI into clinical practice requires an understanding of its limitations and potential, as well as continuous efforts to improve its applications in the medical context.

All authors declare no potential conflict of interest related to this article.

CONTRIBUTIONS OF THE AUTHORS: Each author contributed individually and significantly to the development of this article. LSRO: Conceptualization, Methodology, data acquisition and Writing the paper; RS: Substantial contribution to design, formal analysis and data acquisition; AC: Substantial contribution in the design, interpretation of data, Acquisition of funding; BM: Data curation, analysis and acquisition; WO: Methodology and data acquisition; GF: data interpretation and critical review of its intellectual content; GAVP: Substantial contribution to the design and interpretation of the data; Gabrielle do Amaral Virginio Pereira: Substantial contribution in the design, interpretation of data; LRR: Validation and writing - revision and editing, Final approval of the version of the manuscript to be published.

REFERENCES

1. Sohail SS. A Promising Start and Not a Panacea: ChatGPT's Early Impact and Potential in Medical Science and Biomedical Engineering Research. Ann Biomed Eng. 2024;52(5):1131-1135. doi: 10.1007/s10439-023-03335-6.
2. Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. IBM J Res Dev. 1959;3(3):210–29. doi: 10.1147/rd.33.0210.
3. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, et al. Deep Learning in Medical Imaging: General Overview. Korean J Radiol. 2017;18(4):570-584. doi: 10.3348/kjr.2017.18.4.570.
4. King BF Jr. Artificial Intelligence and Radiology: What Will the Future Hold? J Am Coll Radiol. 2018;15(3 Pt B):501-503. doi: 10.1016/j.jacr.2017.11.017.
5. Pesapane F, Volonté C, Codari M, Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights Imaging. 2018;9(5):745-753. doi: 10.1007/s13244-018-0645-y.
6. Zhang N, Sun Z, Xie Y, Wu H, Li C. The latest version ChatGPT powered by GPT-4o: what will it bring to the medical field? Int J Surg. 2024;110(9):6018-6019. doi: 10.1097/JS9.0000000000001754.
7. Bečulić H, Begagić E, Skomorac R, Mašović A, Selimović E, Pojskić M. ChatGPT's contributions to the evolution of neurosurgical practice and education: a systematic review of benefits, concerns and limitations. Med Glas (Zenica). 2024;21(1). doi: 10.17392/1661-23.
8. He Y, Tang H, Wang D, Gu S, Ni G, Wu H. Will ChatGPT/GPT-4 be a Lighthouse to Guide Spinal Surgeons? Ann Biomed Eng. 2023;51(7):1362-1365. doi: 10.1007/s10439-023-03206-0.
9. The jamovi project. jamovi. [Internet]. 2021. Retrieved from: https://www.jamovi.org/.
10. R Core Team. A Language and environment for statistical computing. [Internet]. 2021. Retrieved from: https://search.gesis.org/publication/zis-RCoreTeam.2021R.
11. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312. doi: 10.2196/45312.
12. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature. 2023;614(7947):214-216. doi: 10.1038/d41586-023-00340-6.
13. Bashshur R, Doarn CR, Frenk JM, Kvedar JC, Woolliscroft JO. Telemedicine and the COVID-19 Pandemic, Lessons for the Future. Telemed J E Health. 2020;26(5):571-573. doi: 10.1089/tmj.2020.29040.rb.
14. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis. 2023;23(4):405-406. doi: 10.1016/S1473-3099(23)00113-5.
15. Ouanes K, Farhah N. Effectiveness of Artificial Intelligence (AI) in Clinical Decision Support Systems and Care Delivery. J Med Syst. 2024;48(1):74. doi: 10.1007/s10916-024-02098-4.
16. Lüscher TF, Wenzl FA, D'Ascenzo F, Friedman PA, Antoniades C. Artificial intelligence in cardiovascular medicine: clinical applications. Eur Heart J. 2024;45(40):4291-4304. doi: 10.1093/eurheartj/ehae465.
17. Han C, Pan Y, Liu C, Yang X, Li J, Wang K, et al. Assessing the decision quality of artificial intelligence and oncologists of different experience in different regions in breast cancer treatment. Front Oncol. 2023;13:1152013. doi: 10.3389/fonc.2023.1152013.
18. Senders JT, Zaki MM, Karhade A V., Chang B, Gormley WB, Broekman ML, et al. An introduction and overview of machine learning in neurosurgical care. Acta Neurochir (Wien). 2018;160(1):29-38. doi: 10.1007/s00701-017-3385-8.
19. Lenke LG, Edwards CC 2nd, Bridwell KH. The Lenke classification of adolescent idiopathic scoliosis: how it organizes curve patterns as a template to perform selective fusions of the spine. Spine (Phila Pa 1976). 2003;28(20):S199-207. doi: 10.1097/01.BRS.0000092216.16155.33.
20. Hoashi JS, Cahill PJ, Bennett JT, Samdani AF. Adolescent scoliosis classification and treatment. Neurosurg Clin N Am. 2013;24(2):173-83. doi: 10.1016/j.nec.2012.12.006.
21. Zhou S, Zhou F, Sun Y, Chen X, Diao Y, Zhao Y, et al. The application of artificial intelligence in spine surgery. Front Surg. 2022;9:885599. doi: 10.3389/fsurg.2022.885599.
22. Yang Q, Steinfeld A, Zimmerman J. Unremarkable AI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM. 2019;(238):1–11. doi. org/10.1145/3290605.3300468
23. Howard A, Wright JG, Hedden D. A comparative study of TLSO, Charleston, and Milwaukee braces for idiopathic scoliosis. Spine (Phila Pa 1976). 1998;23(22):2404-11. doi: 10.1097/00007632-199811150-00009.
24. Janicki JA, Poe-Kochert C, Armstrong DG, Thompson GH. A comparison of the thoracolumbosacral orthoses and providence orthosis in the treatment of adolescent idiopathic scoliosis: results using the new SRS inclusion and assessment criteria for bracing studies. J Pediatr Orthop. 2007;27(4):369-74. doi: 10.1097/01.bpb.0000271331.71857.9a.
25. Tekkeşin Aİ. Artificial Intelligence in Healthcare: Past, Present and Future. Anatol J Cardiol. 2019;22(Suppl 2):8-9. doi: 10.14744/AnatolJCardiol.2019.28661.