Revista Española de Cirugía Ortopédica y Traumatología xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Revista Española de Cirugía Ortopédica y Traumatología

journal homepage: www.elsevier.es/rot

Original

ChatGPT frente a un examen teórico de Cirugía Ortopédica y Traumatología: valor clínico y educativo

ChatGPT in a theoretical examination of Orthopaedic Surgery and Traumatology: clinical and educational value

O. Pujol [a,b,*], M. Guzmán [a], C. Álvaro [c], J. Leal [b,d], J. Minguell [a] y N. Joshi [a]

[a] *Knee Surgery Unit, Orthopaedic Surgery Department, Vall d'Hebron University Hospital, Universitat Autònoma de Barcelona (Departament de Cirurgia), Barcelona, España*
[b] *Knee Surgery Unit, iMove Traumatology, Barcelona, España*
[c] *Sant Joan de Deu University Hospital, Barcelona, España*
[d] *Knee Surgery Unit, Orthopaedic Surgery Department, Hospital Sant Joan de Déu de Manresa - Fundació Althaia, Universitat de Vic, Manresa, Barcelona, España*

INFORMACIÓN DEL ARTÍCULO

Palabras clave:
Inteligencia artificial
Aprendizaje automático
ChatGPT
Cirugía ortopédica y traumatología, Educación
Imágenes médicas

RESUMEN

Introducción: ChatGPT, un *chatbot* de inteligencia artificial (IA) generativa, es una herramienta prometedora de apoyo en el diagnóstico, en la toma de decisiones y en la educación en Cirugía Ortopédica y Traumatología (COT). El objetivo principal de este estudio es evaluar la capacidad de ChatGPT-4o para responder las preguntas de un examen teórico dirigido a residentes de COT. El objetivo secundario es comparar la tasa de aciertos y el patrón de respuestas obtenidos por este *chatbot* con los de los residentes, categorizados según sus años de experiencia.
Métodos: Es un estudio observacional retrospectivo. Se ha analizado el examen teórico de COT realizado por residentes de un hospital terciario español en 2024. El examen incluyó 48 preguntas tipo test (10 con imagen), distribuidas entre distintas subespecialidades. Se registraron las respuestas de ChatGPT-4o y de los residentes para comparar las tasas de aciertos. Además, se analizó la capacidad de acertar preguntas en función de la temática y de la vinculación a imagen.
Resultados: ChatGPT-4o respondió correctamente 34 de 48 preguntas (71%). La tasa de respuestas correctas de ChatGPT fue superior a la media de los residentes de COT (67%), obteniendo una puntuación similar a la de los residentes de quinto año (70%). Sin embargo, mostró una tasa de acierto mucho menor en las preguntas vinculadas a imagen clínica o radiológica (30%).
Conclusiones: ChatGPT-4o es capaz de responder preguntas de un examen teórico sobre COT, obteniendo una tasa de aciertos superior a la media de los residentes de COT y similar a la de los residentes de quinto año. No obstante, la tasa de errores fue del 29,2% y destacó una capacidad más limitada para acertar preguntas vinculadas a imágenes, así como cuestiones que exijan razonamiento clínico complejo. El uso de este modelo de IA no puede sustituir la experiencia y el razonamiento del profesional médico.

Keywords:
Artificial intelligence
Machine learning
ChatGPT
Orthopedic surgery and traumatology
Education
Medical images

ABSTRACT

Introduction: ChatGPT, a generative artificial intelligence (AI) chatbot, represents a potential tool to support diagnosis, decision-making, and education in Orthopaedic Surgery and Traumatology (OST). The primary aim of this study was to evaluate the ability of ChatGPT-4o to answer questions from a theoretical exam designed for OST residents. The secondary aim was to compare the chatbot's score and response patterns with those of residents, stratified by years of training.

* Autor para correspondencia.
Correos electrónicos: oriolp-6@hotmail.com, Oriol.PujolA@autonoma.cat (O. Pujol).

https://doi.org/10.1016/j.recot.2025.10.001
Recibido el 28 de agosto de 2025; Aceptado el 13 de octubre de 2025
Disponible en Internet el xxx

Cómo citar este artículo: O. Pujol, M. Guzmán, C. Álvaro et al., ChatGPT frente a un examen teórico de Cirugía Ortopédica y Traumatología: valor clínico y educativo, Revista Española de Cirugía Ortopédica y Traumatología, https://doi.org/10.1016/j.recot.2025.10.001

Methods: This was a retrospective observational study. A theoretical OST exam administered in 2024 to residents at a Spanish tertiary hospital was analyzed. The exam comprised 48 multiple-choice questions (10 including images) across different subspecialties. The responses of ChatGPT-4o and the residents were recorded to compare accuracy rates. In addition, the ability to correctly answer questions was analyzed according to topic and association with images.
Results: ChatGPT-4o correctly answered 34 out of 48 questions (71%). Its accuracy rate was higher than the average of OST residents (67%), achieving a score comparable to fifth-year residents (70%). However, its performance was notably lower in image-based clinical or radiological questions (30% accuracy).
Conclusion: ChatGPT-4o is capable of answering questions from a theoretical OST examination, achieving a score higher than the average of OST residents and comparable to that of the most experienced residents (fifth-year). However, the error rate was 29.2%, with a notably lower accuracy in questions involving images and those requiring complex clinical reasoning. The use of this AI model cannot replace the expertise and reasoning of medical professionals.

Introducción

Durante la última década se ha producido un desarrollo muy significativo en inteligencia artificial (IA) y aprendizaje profundo (*deep learning* [DL]). Esta tecnología ha transformado la manera en que los humanos abordamos una amplia variedad de tareas, afectando también al ámbito científico y médico[1]. En particular, la aplicación de la IA ha ganado notable relevancia en el análisis de datos clínicos[2], ya que la IA es especialmente eficaz en identificar patrones y realizar clasificaciones binarias[3]. La IA se ha desarrollado también en el área de la Cirugía Ortopédica y Traumatología (COT), donde su uso se está diversificando rápidamente. Se ha incorporado en la educación del paciente y del profesional, así como en la formación específica quirúrgica[4]. También ha incrementado su uso en investigación científica COT[5,6]. Finalmente, puede intervenir en distintas fases del proceso asistencial, como la planificación preoperatoria[7,8] o la predicción de resultados clínicos[9].

En este contexto, ha surgido ChatGPT (OpenAI®, San Francisco, EE.UU.), un *chatbot* de IA generativa que emplea modelos de lenguaje de gran escala (*large language models* [LLM]). Ha demostrado capacidades avanzadas en el procesamiento y la generación de texto en lenguaje natural. Entrenado con grandes volúmenes de datos, este modelo es capaz de interpretar preguntas complejas, analizar de manera contextualizada y generar respuestas argumentadas[10], lo que lo posiciona como una herramienta prometedora en entornos educativos y clínicos.

Diversos estudios han evaluado los conocimientos médicos de ChatGPT mediante la capacidad para responder preguntas de examen en distintas especialidades médicas, incluyendo el ámbito de COT. Se han reportado tasas de acierto que varían entre el 35.8% y el 76%, según el país, la versión del modelo utilizada y la complejidad de las preguntas[11-14].

El objetivo principal de este estudio es evaluar la capacidad de ChatGPT (versión 4o) para responder las preguntas de un examen teórico dirigido a residentes de COT en un hospital de tercer nivel en España. El objetivo secundario es comparar la tasa de aciertos y el patrón de respuestas obtenidos por este *chatbot* con los de los residentes, categorizados según sus años de experiencia.

Métodos

Diseño del estudio

En este estudio comparativo se ha evaluado la capacidad de un modelo de IA generativa (ChatGPT) para responder preguntas de un examen teórico sobre COT. Para el desarrollo de esta investigación se han seguido las recomendaciones de la guía «TRIPOD + AI». Se ha tomado como referencia el examen teórico realizado previamente (2024) a los residentes de COT de un hospital de tercer nivel español. Dicho examen se realiza cada año en el hospital para evaluar el nivel de competencia teórica en COT de los residentes del servicio. El examen es diseñado internamente por los especialistas del servicio de COT. La residencia consta de cinco años, habiendo seis residentes cada año. En el año 2024, este examen escrito contó con 48 preguntas tipo test, con cuatro opciones de respuesta donde solo una era correcta y los errores no restaban puntuación. Diez de las preguntas estaban vinculadas a una imagen clínica o radiológica. Las preguntas estaban divididas equitativamente entre las siguientes subespecialidades: reconstrucción osteoarticular séptica, COT-pediátrica, cirugía de rodilla, cirugía de tobillo-pie, cirugía de antebrazo-mano, cirugía de cadera, reconstrucción osteoarticular oncológica, cirugía de hombro, traumatología y cirugía de raquis.

Se recogieron las respuestas, la tasa de acierto y las puntuaciones de los residentes COT mediante la base de datos electrónica prospectiva del Servicio COT. Se analizó la tasa de acierto en función del año de experiencia de los residentes, la temática de la pregunta y la vinculación o no a imagen. Estos resultados se compararon con los proporcionados por ChatGPT.

Modelo de IA

Se usó la versión más reciente del modelo ChatGPT (Chat Generative Pre-trained Transformer; OpenAI®, San Francisco, EE.UU.) disponible en el momento de realizar el estudio: ChatGPT-4o. Se usó el siguiente *prompt* para optimizar la capacidad de respuesta del sistema de IA: «Tu tarea consiste en responder preguntas sobre cirugía ortopédica y traumatología. Yo proporcionaré las preguntas. Se trata de preguntas tipo test, donde solo una opción es correcta. Debes elegir la opción correcta. Quiero que actúes como un cirujano ortopédico experto. Debes basarte en todos los conocimientos disponibles sobre cirugía ortopédica y traumatología, así como los estándares de manejo establecidos como referencia actual y estar actualizado con los hallazgos de investigación más recientes en el campo. Si no estás seguro de una respuesta, utiliza tus herramientas para obtener información relevante. Debes analizar cuidadosamente cada respuesta antes de responder». Cada pregunta se realizó una única vez. Después de cada respuesta del *chatbot* se abrió un nuevo chat para realizar la siguiente pregunta, con el objetivo de garantizar la independencia de las respuestas entre sí. Las preguntas se realizaron en castellano o catalán, copiando el formato y el idioma original de cada pregunta.

Análisis estadístico

Las variables categóricas se describieron mediante sus valores absolutos, fracciones y porcentajes. Las variables continuas se presentaron mediante su media y desviación estándar. Los grupos se compararon utilizando la prueba Z-test para proporciones. Todos los valores de p fueron bilaterales. Se consideró estadísticamente significativo un valor de $p < 0,05$. El análisis estadístico se realizó utilizando IBM SPSS v. 20.0 (IBM Corp., Armonk, NY, EE.UU.).

Tabla 1
Distribución de la tasa de aciertos de las preguntas teóricas de un examen sobre COT realizado por ChatGPT y por residentes de COT. Se analizan los resultados en función de la subespecialidad temática y de la presencia de imagen en la pregunta

	Tasa de acierto ChatGPT (%)	Tasa de acierto residentes COT (%)
Total (48 preguntas)	70,8	67,4 ± 24
Vinculación a imagen		
No (38/48, 79,2%)	81,6	66,4 ± 26
Sí (10/48, 20,8%)	30,0	71,1 ± 15
Subespecialidad		
Reconstrucción osteoarticular séptica (5/48, 10,4%)	100,0	88,8 ± 11
COT-Pediátrica (5/48, 10,4%)	60,0	59,2 ± 26
Cirugía de rodilla (5/48, 10,4%)	80,0	73,2 ± 27
Cirugía de tobillo-pie (5/48, 10,4%)	40,0	47,6 ± 2
Cirugía de antebrazo-mano (5/48, 10,4%)	60,0	82,0 ± 15
Cirugía de cadera (5/48, 10,4%)	60,0	77,4 ± 10
Reconstrucción osteoarticular oncológica (5/48, 10,4%)	100,0	62,2 ± 13
Cirugía de hombro (4/48, 8,3%)	80,0	60,2 ± 23
Traumatología (5/48, 10,4%)	20,0	50,8 ± 35
Cirugía de raquis (4/48, 8,3%)	100,0	71,3 ± 29

COT: Cirugía Ortopédica y Traumatología.

Resultados

Modelo de IA

ChatGPT acertó 34 de 48 preguntas, correspondiendo a una tasa de acierto del 71%. La distribución de la tasa de aciertos en función de la temática de las preguntas se encuentra en la tabla 1. Destaca que solo acertó el 30% de las preguntas (3/10) vinculadas a imagen clínica o radiológica, siendo un resultado significativamente inferior a su tasa media de acierto ($p < 0,01$). No hubo diferencias en la tasa de aciertos entre las preguntas redactadas en castellano o en catalán: 23/33 (70%) vs 11/15 (73%); $p = 0,8$.

Residentes de COT

Los residentes de COT acertaron de media el 67,4% de las preguntas (67,4 ± 0,9) (tabla 1). Se observó un incremento progresivo en la media de tasa de acierto en función de los años de experiencia de los residentes (R): R1 (54 ± 6), R2 (56 ± 8), R3 (62 ± 9), R4 (69 ± 6) y R5 (70 ± 3) (fig. 1). No se encontraron diferencias en la tasa de acierto entre las preguntas con o sin imagen: 71,1% vs 66,4% ($p = 0,7$).

Modelo de IA vs residentes de COT

La tasa de aciertos obtenida por ChatGPT en el examen teórico en COT fue superior a la tasa media de los residentes de COT (71% vs 67%). De hecho, la puntuación de este modelo de IA fue parecida a la media obtenida por los residentes de quinto año (71% vs 70%) (fig. 1). Sin embargo, su tasa de acierto fue muy inferior en las preguntas vinculadas a imagen (30% vs 71%) (tabla 1).

Discusión

El resultado más relevante de este estudio es que ChatGPT-4o presentó un 29,2% de errores al responder preguntas sobre COT en un examen teórico dirigido a residentes. Destaca también que mostró una tasa de acierto significativamente menor en las cuestiones vinculadas a imagen clínica o radiológica (30%). No obstante, la tasa de respuestas correctas de ChatGPT fue superior a la tasa media de los residentes de COT (71% vs 67%), obteniendo una puntuación similar a la de los residentes de quinto año (71% vs 70%).

ChatGPT es un modelo de IA basado en LLM que ha sido entrenado con una enorme cantidad de texto generado por humanos, incluyendo publicaciones científicas[15]. Gracias a este entrenamiento, maneja un volumen masivo de datos al tratar cuestiones relacionadas con distintas especialidades médicas. Una de las ventajas más destacadas de ChatGPT es su capacidad para comprender lenguaje natural y generar texto original de forma coherente y contextualizada. Esto permite que el usuario le pueda realizar preguntas y mantener un diálogo lógico. Todos estos factores hacen que ChatGPT sea una herramienta con mucho potencial en el entorno médico[16]. Las últimas versiones de este sistema de IA también permiten integrar, interpretar y generar imágenes. Sin embargo, estudios previos han evidenciado que tareas aparentemente sencillas, como el reconocimiento de imágenes u objetos, pueden suponer un reto considerable para los modelos de IA, ya que involucran procesos cognitivos abstractos que resultan difíciles de codificar y de replicar en sistemas basados en lenguaje natural como los *chatbots*[17,18]. El potencial en el análisis de imágenes de ChatGPT aún no ha sido plenamente desarrollado[19]. De hecho, en nuestro estudio destaca que ChatGPT presentó una tasa de acierto muy inferior en las preguntas vinculadas a imagen (30% vs 71%). Adicionalmente, se observó que podía confundir zonas de la imagen de escasa definición con un trazo de fractura, y por este motivo es recomendable utilizar documentos con buena calidad visual al valorar imágenes radiológicas por ChatGPT.

Nuestros resultados son comparables a los obtenidos por la versión ChatGPT-4o al realizar exámenes sobre la especialidad de COT en otros países. Pamuk et al. observaron una tasa de acierto del 76% en el *Turkish Orthopedics and Traumatology Board Examination*, obteniendo una nota superior al 98,7% de los candidatos humanos[11]. Maraqa et al. reportaron una puntuación del 74,8% en el *French Orthopedic and Trauma Surgery Exam*, superior a la del 70,8% obtenida por los residentes[13]. Estos autores destacaron la alta capacidad de ChatGPT para responder preguntas teóricas en formato texto. Sin embargo, Cuthbert y Simpson evaluaron la capacidad de ChatGPT para superar el examen de *Fellowship* del *Royal College of Surgeons* en COT y observaron que el modelo alcanzó únicamente un 35,8% de respuestas correctas[14]. Esta puntuación fue significativamente inferior al umbral de aprobación del examen y a la media obtenida por los candidatos humanos. Entre las principales limitaciones identificadas en ChatGPT se encontraron su incapacidad para ejercer juicio clínico y la falta de razonamiento lógico requerido para resolver preguntas que demandan un procesamiento cognitivo más complejo. Esta debilidad también se evidenció en nuestro estudio. Observamos que ChatGPT presenta una elevada capacidad para resolver preguntas teóricas directas, ya que es capaz de obtener la información a través del volumen inmenso de datos con los que se ha entrenado. Sin embargo, tiene más dificultades para resolver preguntas relacionadas

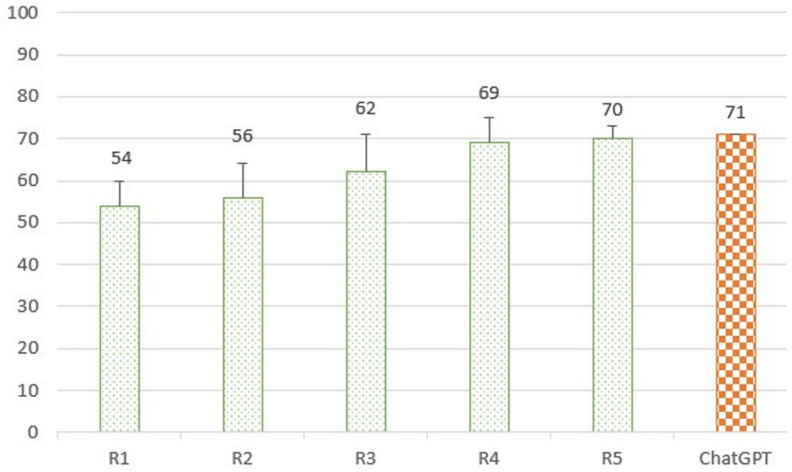

Figura 1. Gráfico comparando la tasa de aciertos (%) obtenida en un examen teórico sobre COT entre el modelo de IA ChatGPT (color naranja, patrón sombreado cuadrícula) y residentes de COT categorizados según sus años de experiencia (color verde, patrón sombreado punteado).
COT: Cirugía Ortopédica y Traumatología; R: residente.

con un caso clínico o que requieran un razonamiento complejo. Por ejemplo, el modelo de IA no supo responder correctamente una pregunta sobre cómo equilibrar la tensión ligamentosa en una prótesis de rodilla con una discrepancia entre la laxitud en flexión-extensión (sin embargo, esta pregunta fue acertada por el 65% de residentes). Por otro lado, también detectamos la dificultad de ChatGPT para entender conceptos más inespecíficos como «relevante» o «probable», más fácilmente comprensibles para los candidatos humanos. Por todo esto, es probable que las diferencias observadas en las tasas de acierto entre subespecialidades (tabla 1) no se explican necesariamente por el nivel de conocimiento en cada área, sino más posiblemente por la naturaleza y el enfoque de las preguntas formuladas, así como por la presencia de cuestiones vinculadas a imagen. Estos resultados refuerzan la necesidad de evaluar de manera crítica la fiabilidad y la aplicabilidad de la IA en escenarios clínicos reales, especialmente en aquellos de alta complejidad[20].

Existen múltiples sistemas de IA generativa basados en lenguaje natural, como Claude, GROK, Gemini o DeepSeek, que presentan sus particularidades, sus pros y sus contras. A pesar de sus claros beneficios, todos los *chatbots* enfrentan aún limitaciones estructurales. 1) Problemas de sesgo y generalización, ya que estos modelos reproducen las limitaciones presentes en sus datos de entrenamiento. Por ejemplo, la infrarrepresentación de ciertas subpoblaciones en bases de datos clínicas puede generar recomendaciones diagnósticas o terapéuticas inadecuadas para estos pacientes[21,22]. 2) Falta de empatía y juicio clínico adaptativo, limitando su utilidad en contextos médicos complejos[23]. 3) Problemas relacionados con la privacidad y la propiedad de los datos clínicos de los pacientes[24]. Por ejemplo, al plantear preguntas al *chatbot* usando documentos o texto que contengan datos personales. 4) Falta de claridad en la asignación de responsabilidades, que plantea riesgos legales frente a errores cometidos al seguir las recomendaciones de la IA[25]. 5) Controversias éticas[26]. 6) Falta de explicabilidad; el modelo procesa los datos de entrada y genera una salida de forma opaca, sin proporcionar información sobre el razonamiento subyacente (modelos de «caja negra»)[27]. 7) Falta de confianza en el sistema de IA y miedo a la pérdida de capacidades humanas[27]. 8) Riesgo de alucinaciones, situaciones donde la IA responde con contundencia y de forma convincente, aunque no tenga certeza de la información que proporciona o incluso haya generado una respuesta ficticia[28]. Todas estas limitaciones deben ser detectadas, reconocidas y comunicadas claramente, con el objetivo de que los profesionales clínicos comprendan el alcance y las restricciones actuales de estos sistemas.

Los resultados de este estudio confirman el potencial de la IA en el ámbito de la COT. Sin embargo, a pesar de que ChatGPT obtuvo una puntuación superior a la media de candidatos humanos, destaca que presentó un 29,2% de respuestas erróneas en un examen dirigido a residentes. Este dato es relevante, porque este *chatbot* puede ser una herramienta de consulta habitual para los cirujanos en formación. Por ello, la implementación de la IA debe enmarcarse dentro de una estrategia supervisada y crítica, subrayando la importancia de seguir promoviendo el razonamiento clínico, la humanización de la medicina, la seguridad del paciente y la ética en la formación de los residentes. En estos aspectos, los profesionales humanos continúan siendo insustituibles frente a la IA.

Este estudio presenta limitaciones. El uso de un único modelo de IA (ChatGPT) limita la posibilidad de generalizar los hallazgos a otros modelos. Además, el estudio refleja la capacidad para responder preguntas sobre COT por una versión concreta de ChatGPT (4o), por lo que futuras actualizaciones del modelo podrían modificar su eficacia. El *prompt* concreto usado al realizar las preguntas y el número de iteraciones también podría influir en los resultados. Por otro lado, el examen se llevó a cabo por residentes de un único centro. Aun así, en el hospital del estudio se forman seis residentes cada año, aportando un número considerable de candidatos humanos. Aunque se trata de un examen no oficial, este test se realiza cada año y sigue un formato y una estructura similares a los de otras pruebas formales. La heterogeneidad en la redacción de las preguntas debido a que han sido escritas por diferentes profesionales también puede haber afectado a la capacidad de ChatGPT para responder correctamente. Finalmente, el tamaño muestral y el número de preguntas relativamente pequeños podría limitar el poder estadístico del estudio.

Son fortalezas del estudio el haber evaluado tanto preguntas en formato texto como vinculadas a imagen, ofreciendo un análisis más completo sobre las capacidades de este modelo de IA. Además, se ha comparado la tasa de aciertos y patrón de respuestas obtenidos por el *chatbot* con los de los residentes, considerando también sus años de experiencia. Finalmente, se ha usado un *prompt* exhaustivo (reflejado en el apartado métodos) para optimizar la capacidad de respuesta del sistema.

Conclusiones

ChatGPT-4o es capaz de responder preguntas de un examen teórico sobre COT, obteniendo una tasa de aciertos superior a la media de los residentes de COT y similar a la de los residentes de quinto año. Este hallazgo refuerza su potencial como herramienta de apoyo y formación dentro del ámbito médico. No obstante, la tasa de errores fue del 29,2% y destacó una capacidad más limitada para acertar preguntas vinculadas a imágenes clínicas o radiológicas, así como cuestiones que exijan razonamiento clínico complejo. El uso de este modelo de IA no puede sustituir la experiencia y el razonamiento del profesional médico; debe enmar-

carse dentro de una estrategia de implementación crítica y supervisada, garantizando la seguridad del paciente y la responsabilidad ética.

Nivel de evidencia

Nivel de evidencia IV.

Financiación

Esta investigación no recibió ninguna subvención específica de organismos de financiación de los sectores público, comercial o sin ánimo de lucro.

Consideraciones éticas

El estudio estuvo exento de aprobación ética, y se le concedió la exención de exigir el consentimiento por escrito de los pacientes.

Contribución de los autores

Todos los autores contribuyeron por igual a este trabajo. Todos los autores contribuyeron en la concepción y el diseño del estudio, la preparación del material, la recopilación y el análisis de datos. El primer borrador del manuscrito fue escrito por OP, y todos los autores comentaron las versiones del manuscrito. Todos los autores leyeron y aprobaron el manuscrito final.

Conflicto de intereses

Ninguno.

Bibliografía

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11:887, http://dx.doi.org/10.3390/healthcare11060887.
2. Yates EJ, Yates LC, Harvey H. Machine learning «red dot»: open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. *Clin Radiol*. 2018;73:827–831, http://dx.doi.org/10.1016/j.crad.2018.05.015.
3. Shin H-C, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*. 2016;35:1285–1298, http://dx.doi.org/10.1109/TMI.2016.2528162.
4. St Mart JP, Goh EL, Liew I, Shah Z, Sinha J. Artificial intelligence in orthopaedics surgery: transforming technological innovation in patient care and surgical training. *Postgrad Med J*. 2023;99:687–694, http://dx.doi.org/10.1136/postgradmedj-2022-141596.
5. Giorgino R, Alessandri-Bonetti M, Luca A, et al. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. *Front Surg*. 2023;10:1284015, http://dx.doi.org/10.3389/fsurg.2023.1284015.
6. Guan J, Li Z, Sheng S, et al. An artificial intelligence-driven revolution in orthopedic surgery and sports medicine. *Int J Surg*. 2025;111:2162–2181, http://dx.doi.org/10.1097/JS9.0000000000002187.
7. Lambrechts A, Wirix-Speetjens R, Maes F, van Huffel S. Artificial intelligence based patient-specific preoperative planning algorithm for total knee arthroplasty. *Front Robot IA*. 2022;9:840282, http://dx.doi.org/10.3389/frobt.2022.840282.
8. Khaje Mozafari J, Ali Moshtaghioon S, Mani Mahdavi S, Ghaznavi A, Behjat M, Yeganeh A. The role of artificial intelligence in preoperative planning for total hip arthroplasty: a systematic review. *Front Artif Intell*. 2024;7:1417729, http://dx.doi.org/10.3389/frai.2024.1417729.
9. Vikas K, Christopher R, Steven O, et al. Using machine learning to predict clinical outcomes after shoulder arthroplasty with a minimal feature set. *J Shoulder Elbow Surg*. 2021;30:225–236, http://dx.doi.org/10.1016/j.jse.2020.07.042.
10. Taylor WL4th, Cheng R, Weinblatt A, Bergstein V, Long WJ. An artificial intelligence chatbot is an accurate and useful online patient resource prior to total knee arthroplasty. *J Arthroplasty*. 2024;39:358–362, http://dx.doi.org/10.1016/j.arth.2024.02.005.
11. Pamuk Ç, Uyanik AF, Kuyucu E, Uğurlar M. Can ChatGPT pass the Turkish Orthopedics and Traumatology Board Examination? Turkish orthopedic surgeons versus artificial intelligence. *Ulus Travma Acil Cerrahi Derg*. 2025;31:310–315, http://dx.doi.org/10.14744/tjtes.2025.07724.
12. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery Examination? Orthopaedic residents versus ChatGPT. *Clin Orthop*. 2023;481:1623–1630, http://dx.doi.org/10.1097/CORR.0000000000002704.
13. Maraqa N, Samargandi R, Poichotte A, Berhouet J, Benhenneda R. Comparing performances of French orthopaedic surgery residents with the artificial intelligence ChatGPT-4/4o in the French diploma exams of orthopaedic and trauma surgery. *Orthop Traumatol Surg Res*. 2024, http://dx.doi.org/10.1016/j.otsr. 2024.104080.
14. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: Can Chat Generative Pre-trained Transformer (ChatGPT) pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? *Postgrad Med J*. 2023;99:1110–1114, http://dx.doi.org/10.1093/postmj/qgad053.
15. Gilson A, Safranek CW, Huang T, 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, http://dx.doi.org/10.2196/45312.
16. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. *Croat Med J*. 2023;64:1–3, http://dx.doi.org/10.3325/cmj.2023.64.1.
17. Khoriati A, Shahid Z, Fok M, et al. Artificial intelligence and the orthopaedic surgeon: a review of the literature and potential applications for future practice: current concepts. *J ISAKOS*. 2024;9:227–233, http://dx.doi.org/10.1016/j.jisako.2023.10.015.
18. Zhu J, Jiang Y, Chen D, et al. High identification and positive-negative discrimination but limited detailed grading accuracy of ChatGPT-4o in knee osteoarthritis radiographs. *Knee Surg Sports Traumatol Arthrosc*. 2025;33:1911–1919, http://dx.doi.org/10.1002/ksa.12639.
19. Hüseyin Temel M, Erden Y, Bağcıer F. Evaluating artificial intelligence performance in medical image analysis: Sensitivity, specificity, accuracy, and precision of ChatGPT-4o on Kellgren-Lawrence grading of knee X-ray radiographs. *Knee*. 2025;55:79–84, http://dx.doi.org/10.1016/j.knee.2025.04.008.
20. Sweed T, Mabrouk A, Dawson M. Transforming orthopaedics with AI: insights from a custom ChatGPT on ESSKA osteotomy consensus. *Knee Surg Sports Traumatol Arthrosc*. 2025;33:1557–1559, http://dx.doi.org/10.1002/ksa.12653.
21. Kline A, Wang H, Li Y, et al. Multimodal machine learning in precision health: a scoping review. *Nat Portf*. 2022;5:171, http://dx.doi.org/10.1038/s41746-022-00712-8.
22. Kayaalp ME, Prill R, Sezgin EA, Cong T, Królikowska A, Hirschmann MT. DeepSeek versus ChatGPT: multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation — Redefining innovation in research and practice. *Knee Surg Sports Traumatol Arthrosc*. 2025;33:1553–1556, http://dx.doi.org/10.1002/ksa.12628.
23. Astărăstoae V, Rogozea LM, Leaşu F, Ioan BG. Ethical dilemmas of using artificial intelligence in medicine. *Am J Ther*. 2024;31:388–397, http://dx.doi.org/10.1097/MJT.0000000000001693.
24. Adams TL, Leslie K, Myles S, Moraes B. Regulating professional ethics in a context of technological change. *BMC Med Ethics*. 2024;25:143, http://dx.doi.org/10.1186/s12910-024-01140-x.
25. Sendak MP, Liu VX, Beecy A, et al. Strengthening the use of artificial intelligence within healthcare delivery organizations: balancing regulatory compliance and patient safety. *J Am Med Inform Assoc*. 2024;31:1622–1627, http://dx.doi.org/10.1093/jamia/ocae119.
26. Hakam HT, Prill R, Korte L, et al. Human-written vs AI-generated texts in orthopedic academic literature: comparative qualitative analysis. *JMIR Form Res*. 2024;8:e52164, http://dx.doi.org/10.2196/52164.
27. Oettl FC, Oeding JF, Samuelsson K. Explainable artificial intelligence in orthopedic surgery. *J Exp Ortop*. 2024;11:e12103, http://dx.doi.org/10.1002/jeo2.12103.
28. Sun Y, Sheng D, Zhou Z, Wu Y. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanit Soc Sci Commun Vol*. 2024;11:1278, http://dx.doi.org/10.1057/s41599-024-03811-x.