¿Influye la especialidad médica en las respuestas de ChatGPT? Un análisis sobre el diagnóstico del dolor lumbar

Annika Nack[1], Xabier Michelena[2], Pol Maymó-Paituvi[3], Cristina Calomarde-Gómez[1], David Lobo-Prat[4], Asier García-Alija[5], Raquel Ugena-García[1], Maria Aparicio[1], Paola Vidal-Montal[3], Diego Benavent[3]

[1] Servei de Reumatologia, Hospital Universitari Germans Trias i Pujol; [2] Servei Català de Salut i Hospital Universitari de la Vall d'Hebron; [3] Hospital Universitari de Bellvitge; [4] Hospital Universitari Doctor Josep Trueta; [5] Hospital Universitari de la Santa Creu i Sant Pau

Introducción

El **dolor lumbar (DL)** es una causa frecuente de consulta y puede tener orígenes **reumatológicos, mecánicos, neurológicos o sistémicos**. Su evaluación requiere, a menudo, la participación de **varias especialidades médicas**.

La **inteligencia artificial (IA)**, en particular herramientas de **procesamiento de lenguaje natural** tiene un creciente potencial para **apoyar el diagnóstico diferencial** en este contexto.

Sin embargo, se desconoce si su rendimiento puede **verse influido por el encuadre clínico** (es decir, la especialidad médica simulada en la consulta).

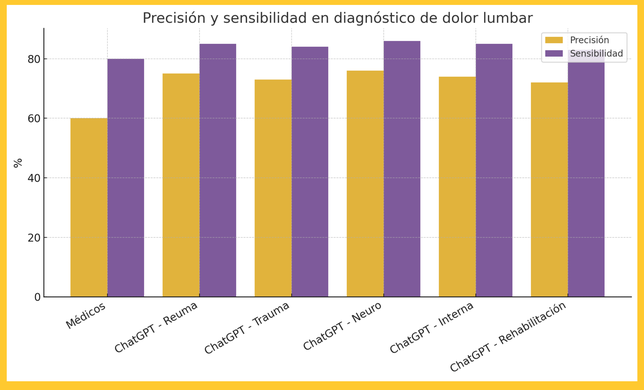
Objetivos

- Evaluar si las **respuestas de ChatGPT varían según la especialidad médica simulada** (reumatología, neurología, traumatología, medicina interna o rehabilitación) en el abordaje del **dolor lumbar.**
- Comparar la **precisión y sensibilidad** diagnóstica de ChatGPT **con la de especialistas en reumatología** ante casos clínicos simulados.

Métodos

- Se escogieron **10 casos clínicos reales** de DL, extraídos de **exámenes oficiales para reumatología en España** (5 casos de DL de causa reumatológica y 5 no reumatológica).
- Participaron **10 médicos especialistes en enfmerdades reumatológicas y musculoesqueléticas**, con ≥5 años de experiencia.
- Cada caso clínico fue respondido por los médicos y por ChatGPT-4o, generando **tres opciones diagnósticas** por participante. En el caso de **ChatGPT,** se le pidió actuar **como si fuera un especialista en reumatología, neurología, medicina interna, rehabilitación o traumatología**, evaluando el caso desde cada uno de estos enfoques.
- Se compararon las respuestas con el **diagnóstico oficial** del examen (estándar de referencia), evaluando:
  - **Precisión**: acierto en el diagnóstico principal
  - **Sensibilidad**: inclusión del diagnóstico correcto entre los tres primeros
  - **Tiempo** de respuesta en contestar todas las preguntas por parte de los médicos y de ChatGPT

ChatGPT acierta en los diagnósticos clínicos independientemente de la especialidad médica indicada en las instrucciones.



ChatGPT alcanza mayor precisión y sensibilidad diagnóstica que los médicos, independientemente de la especialidad indicada.

Resultados

- Entre los médicos y ChatGPT se generaron un total de **528 diagnósticos**, agrupados en **39 categorías clínicas**.
- En la mayoría de las respuestas generadas por ChatGPT, el diagnóstico correcto se encontraba en las primeras posiciones con una **precisión diagnóstica de ChatGPT** osciló entre el **70 % y 80 %** según la especialidad simulada. La **sensibilidad**, entre el **80 % y 90 %**.
- No hubo **diferencias significativas entre especialidades simuladas** en precisión ($p = 0,80$) ni en sensibilidad ($p = 0,68$).
- Comparado con los médicos:
  - **ChatGPT fue más preciso**: 75 % vs. 60 % ($p < 0,001$)
  - **Mayor sensibilidad**: 85 % vs. 80 % ($p = 0,02$)
  - **Más rápido**: 2,3 min vs. 12,4 min ($p < 0,01$)

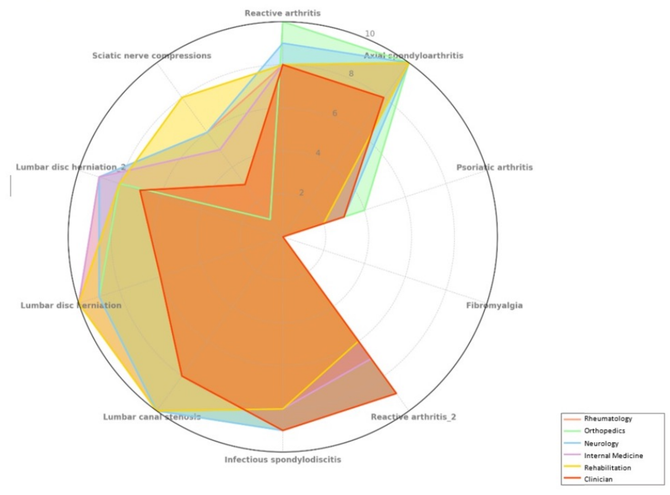


Figura 2. Porcentaje de respuestas correctas para cada participante y cada especialidad simulada. Las líneas de colores representan a ChatGPT simulando distintas especialidades. Su rendimiento diagnóstico fue similar entre ellas, lo que sugiere que la especialidad indicada no influye en la precisión. Solo en el caso correspondiente a fibromialgia no se obtuvo ningún acierto.

Conclusiones

ChatGPT mantiene un rendimiento diagnóstico alto y constante**, sin verse influido por la especialidad médica simulada**.

En esta serie de casos de dolor lumbar, fue **más preciso, más sensible y más rápido** que los médicos participantes.

N.º 203