

# Médicos vs. IA: ¿Quién quieres que te atienda?

Prof. Miguel Lafuente Blasco

## 1. Contexto del estudio

Bajo el nombre MEDICOS\_vs\_GPT, se encuentra un conjunto de datos reales procedente de un estudio publicado en 2025 en *Nature Medicine* por Goh *et al.*<sup>1</sup>. El estudio analiza el impacto del uso de inteligencia artificial (IA) en la toma de decisiones clínicas por parte de médicos.

El objetivo principal es evaluar si la asistencia de un modelo de lenguaje de gran tamaño (LLM, por sus siglas en inglés *Large Language Model*), concretamente GPT-4, mejora el desempeño en tareas clínicas como el razonamiento diagnóstico y el manejo de pacientes.

A continuación, se muestra un extracto del artículo original:

« *We enrolled 92 physicians to participate in the study, which was conducted from 30 November 2023 to 21 April 2024. Participants were randomized evenly between the LLM and conventional resources [...]*

»[...] *From these 92 physicians, 400 cases were scored in total: 176 from the group of physicians using the LLM, 199 from physicians using only conventional resources, and 25 from the LLM alone.* »

El conjunto incluye 400 resoluciones de casos realizadas por 92 médicos. Esto se debe a que cada médico resolvió, por lo general, unos cinco casos diferentes. En cada resolución, se asignó aleatoriamente si el médico contaba con el apoyo de GPT-4 o solo podía usar recursos convencionales. Además, algunos casos fueron resueltos exclusivamente por la IA, sin intervención humana.

La descripción detallada de las variables incluidas en el estudio se encuentra en el Anexo A.

**Fuente:** Goh, E., et al. (2025). *GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial*. *Nature Medicine*. <https://doi.org/10.1038/s41591-024-03456-y>

**Palabras clave:** Análisis bivariante, inferencia básica de comparaciones.

**Número de casos:** 400

**Variables:** 4 categóricas y 9 numéricas (base de datos procesada en español)

---

<sup>1</sup>Goh, E. *et al.* (2025). GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial. *Nature Medicine*. Disponible en: <https://www.nature.com/articles/s41591-024-03456-y>

## 2. Preparación de la base de datos

Abre la base de datos en *jamovi*<sup>2</sup> y revisa el tipo y la codificación de cada variable.

- (a) Asegúrate de que todas las variables estén correctamente clasificadas según su tipo de medida y codificación. Modifica lo que sea necesario.
- (b) Reordena las categorías de las siguientes variables según estos criterios:
  - Aunque **Tipo de Médico** es una variable nominal, ordena sus niveles (sin cambiar el tipo) desde el de mayor rango profesional al de menor.  
*Nota:* Los residentes están en formación y trabajan bajo la supervisión de los médicos adjuntos.
  - En **Grupo de Estudio**, ordena las categorías desde el menor al mayor uso de IA.
  - En **Especialidad**, coloca la categoría *Sin especialidad* en último lugar.
- (c) El tiempo por caso se registra en segundos. Transfórmalo a minutos, crea una nueva variable con ese formato y elimina la original.
- (d) Renombra todas las variables cuyo nombre siga el patrón **Porcentaje X** para que pasen a llamarse **X (%)**. Esto facilitará su identificación en los análisis.
- (e) Calcula la variable **Total (%)** como el porcentaje de puntos obtenidos respecto al máximo posible en cada caso.  
*Nota:* Si el nombre contiene espacios, puedes usar la comilla inversa para que *jamovi* lo reconozca correctamente, por ejemplo: ``Total (%)``.
- (f) Comprueba si hay valores faltantes o atípicos en alguna variable. Presenta una tabla con el número de datos perdidos por variable.
  - a) En **Años de Experiencia** y **Experiencia con GPT**, observarás que hay el mismo número de datos faltantes. Por otro lado, haz un tabla de frecuencias de la variable **Tipo de Médico** y da una explicación coherente de la razón por la que hay esos valores faltantes en esas dos variables.
  - b) Algunos casos no tienen registrada una especialidad. Filtra solo esos casos y comprueba qué tipo de médico son. A partir de esto, argumenta por qué conviene renombrar el nivel *Sin especialidad* y haz el cambio adecuado.
- (g) Mueve la variable **Total (%)** a la primera columna de la base de datos.  
*Nota:* Puedes reorganizar columnas con `ctrl + c`, `ctrl + v` y eliminar la anterior si lo deseas.

Antes de continuar, asegúrate de que la base de datos esté completamente preparada. Si detectas errores o valores incoherentes, probablemente se haya omitido algún paso anterior.

<sup>2</sup>Esta práctica está pensada para resolverse con *jamovi*, pero también puede completarse con cualquier software estadístico básico.

### 3. Resumen descriptivo univariante

Este apartado tiene como objetivo describir las características de los responsables de cada caso evaluado. Como un mismo médico puede haber resuelto varios casos, no se describe a los participantes individualmente, sino a los evaluadores asociados a cada resolución. Es decir, cada fila de la base de datos se analiza como una unidad independiente, aunque algunos médicos aparezcan varias veces.

- (a) Calcula cuántos casos fueron resueltos por médicos humanos y cuántos por la IA sin intervención médica. Expresa estos datos en términos absolutos y porcentuales. Presenta esta información con una tabla y una frase que resuma la comparación.

En los siguientes apartados, céntrate exclusivamente en los casos resueltos por médicos humanos (excluye las resoluciones hechas solo por GPT-4 aplicando un filtro).

- (b) Describe la distribución de la variable **Tipo de Médico**. Presenta un gráfico adecuado, una tabla de frecuencias, y redacta una frase breve que resuma lo observado (¿quién resuelve más casos? ¿hay equilibrio entre grupos?).
- (c) Haz lo mismo con la variable **Especialidad**: presenta un gráfico, una tabla de frecuencias y una frase breve que describa cómo se reparten las especialidades entre los responsables de los casos.
- (d) Analiza la variable **Años de Experiencia**. Presenta gráfico adecuado, indica las medidas más relevantes (media, mediana, desviación estándar, mínimo y máximo) y redacta un párrafo breve que describa cómo es la distribución de la experiencia profesional entre los médicos.
- (e) Evalúa el grado de familiaridad de los médicos con el uso de IA, utilizando la variable **Experiencia con GPT**. Incluye una tabla de frecuencias, un gráfico y una breve interpretación: ¿la mayoría ha usado IA antes o no? ¿con qué frecuencia?
- (f) Describe el desempeño global de los médicos a través de la variable **Total (%)**. Utiliza un gráfico adecuado, proporciona las principales medidas descriptivas (media, mediana, etc.) e incluye una frase resumen sobre cómo ha sido, en general, el rendimiento en los casos clínicos.

### 4. Inferencia Estadística

En cada pregunta de esta sección, se deberá:

1. Mostrar un gráfico adecuado, si es posible, para ilustrar visualmente la distribución de los datos del problema y facilitar la interpretación de los resultados. También se puede incluir una tabla con medidas descriptivas si se considera relevante.
2. Aplicar al menos un test estadístico para responder a la pregunta planteada. Si son necesarias pruebas adicionales para comprobar hipótesis previas o elegir el test adecuado, simplemente presenta los resultados sin justificación adicional (copia directamente la tabla con el resultado del test desde *jamovi*).

3. Finalizar el apartado, siempre que sea posible, con una o dos frases breves en formato de artículo científico<sup>3</sup>, que describan:
- El objetivo del análisis.
  - La prueba estadística utilizada.
  - La estimación de interés (ya sea un parámetro o una diferencia).
  - El p-valor del test y un intervalo de confianza de la estimación, si es posible.

Todas las pruebas estadísticas y los intervalos de confianza se calcularán con un nivel de confianza del 95 %.

#### 4.1. Perfil de los médicos

- (a) ¿Se puede considerar que la proporción de casos resueltos por residentes y adjuntos es la misma?
- (b) ¿Se puede considerar que la proporción de casos resueltos por residentes y adjuntos es la misma en cada especialidad médica? Además, describe el perfil de los médicos que resuelven los casos según su especialidad.
- (c) ¿Los médicos residentes y adjuntos muestran diferencias en su experiencia con la IA?
- (d) ¿Hay evidencia suficiente para argumentar que la experiencia previa con la IA varía según la especialidad del médico?

#### 4.2. Caracterizando las variables objetivo: desempeño y tiempo de resolución

- (a) Muestra en una tabla, para cada grupo de estudio, un intervalo de confianza para la media del desempeño total (Puntuación total (%)).
- (b) ¿Se puede asumir que, en alguno de los grupos de estudio, la puntuación total (en porcentaje) es mayor del 35 %?
- (c) ¿Por qué el intervalo de confianza del grupo *Solo GPT-4* es más amplio que el de *GPT-4 + Médico*?
- (d) Dado que el tiempo de resolución es una variable altamente asimétrica, muestra en una tabla, para cada grupo de estudio, un intervalo de confianza para la mediana del tiempo de resolución.

#### 4.3. Relación entre desempeño y características del médico

- (a) ¿Existe una diferencia significativa en el desempeño final (Puntuación total (%)) entre los distintos niveles de experiencia profesional (*Médico Adjunto* o *Residente*)?

---

<sup>3</sup>En el capítulo de inferencia del siguiente manual se incluyen numerosos ejemplos que puedes adaptar a este contexto: Lafuente Blasco, M. (2025). *Análisis de datos con Jamovi. Guía práctica para la investigación científica* (1ª ed.). Servicio de Publicaciones, Universidad de Zaragoza. Disponible en: <https://zagan.unizar.es/record/151518>

- (b) ¿Se observa una diferencia significativa en el tiempo empleado para la resolución de cada caso según el nivel de experiencia profesional (*Médico Adjunto* o *Residente*)?
- (c) ¿Existen diferencias significativas en el desempeño final (Puntuación total (%)) según el área de especialización?
- (d) ¿Cómo varía, como patrón general, el desempeño final (Puntuación total (%)) en función de la antigüedad de trayectoria profesional (*Años de Experiencia*)? Interpreta el gráfico de dispersión.

#### 4.4. Relación entre experiencia, tiempo empleado y desempeño

- (a) ¿Los profesionales con más años de experiencia son más rápidos en la resolución de casos?
- (b) ¿Existe una relación entre el área de especialización y el tiempo empleado por caso?
- (c) ¿El tiempo empleado en la resolución de cada caso influye en el desempeño final (Puntuación total (%))?

En los siguientes apartados de esta sección, escribe solo una frase mostrando el estadístico adecuado con su intervalo de confianza.

- (d) ¿Los profesionales con más años de experiencia obtienen mejores resultados en cuanto al diagnóstico (*Diagnóstico (%)*)?
- (e) ¿Los profesionales con más años de experiencia hacen un mejor manejo de los casos (*Manejo (%)*)?
- (f) ¿Los profesionales con mayor experiencia recuerdan y aplican con mayor precisión hechos médicos establecidos (*Conocimiento factual (%)*)?
- (g) Interpreta en una única frase los tres apartados anteriores conjuntamente.

#### 4.5. Diferentes facetas del desempeño

- (a) Para los médicos humanos (excluyendo solo a las evaluaciones realizadas por GPT-4), examina si el desempeño en diagnóstico difiere significativamente del desempeño en manejo.
- (b) A partir de las evaluaciones realizadas por GPT4, argumenta si se presentan diferencias significativas en las distintas dimensiones del desempeño (manejo, conocimiento factual y diagnóstico).

#### 4.6. Relación entre el uso de IA y el desempeño

- (a) ¿Existe una diferencia en el desempeño final (Puntuación total (%)) entre los médicos que utilizaron IA y aquellos que solo usaron recursos convencionales?
- (b) Dentro del grupo de médicos que usaron IA, ¿existe una relación entre la experiencia previa con IA y el desempeño final (Puntuación total (%))?

- (c) ¿Existe una diferencia en las medianas del tiempo empleado en la resolución de los casos entre los médicos que usaron IA y los que no? Describe y compara los resultados.
- (d) ¿El tiempo empleado en la resolución de los casos varía entre médicos con apoyo de IA, médicos que usaron solo recursos convencionales y la IA en solitario?
- (e) ¿Se observan diferencias en el desempeño final (**Puntuación total (%)**) entre médicos con apoyo de IA, médicos que usaron solo recursos convencionales y la IA en solitario?

#### 4.7. Comparación del rendimiento entre médicos y médicos con ayuda de la IA

En esta sección se compara la precisión en distintos aspectos de la toma de decisiones clínicas entre los médicos que utilizaron únicamente recursos convencionales y aquellos que emplearon IA.

- (a) Analiza y argumenta cuál de los dos grupos obtuvo un mejor desempeño en la precisión diagnóstica (**Diagnóstico (%)**).
- (b) Explica cuál demostró un mayor dominio de hechos médicos establecidos, basándote en los resultados obtenidos (**Conocimiento factual (%)**).
- (c) Evalúa qué opción tomó decisiones más acertadas en la gestión de los casos y justifica tu respuesta con datos objetivos (**Manejo (%)**).

### 5. Conclusiones

A partir de los resultados obtenidos, elabora una síntesis de las conclusiones más relevantes, argumentando con datos estadísticos, sobre el impacto de la IA en la efectividad y el coste en la toma de decisiones clínicas. Considera tanto la influencia en la precisión de los diagnósticos como la posible reducción en los tiempos de resolución de los casos.

## A. Descripción de las variables

- **Tipo de Médico:** Categoriza a los participantes en función de su nivel de experiencia.
  - *Médico Adjunto:* Profesionales con experiencia en la práctica clínica.
  - *Residente:* Médicos en formación.
  - *Solo GPT-4:* Respuestas generadas sin intervención humana.
- **Grupo de Estudio:** Variable categórica que indica si el médico utilizó GPT-4 o solo recursos convencionales. Categorías:
  - *Solo Médico:* Grupo que solo usó recursos convencionales.
  - *GPT-4 + Médico:* Grupo que utilizó la IA para apoyar la toma de decisiones.
  - *Solo GPT-4:* Respuestas generadas exclusivamente por el modelo de lenguaje sin intervención humana.
- **Especialidad:** Área de la medicina en la que trabaja el médico:
  - *Medicina de Urgencias*
  - *Medicina Familiar*
  - *Medicina Interna*
  - *Sin especialidad especificada*
- **Años de Experiencia:** Años de práctica médica del participante.
- **Experiencia con GPT:** Indica la frecuencia con la que el médico ha usado IA en su práctica profesional. Categorías:
  - *Nunca lo he usado*
  - *Lo usé una vez*
  - *Raramente (menos de 1 vez/mes)*
  - *Ocasionalmente (más de 1 vez/mes pero menos de 1 vez/semana)*
  - *Frecuentemente (1 vez/semana o más)*
- **Puntuación Total:** Puntaje obtenido por cada médico en la prueba de razonamiento clínico.
- **Puntuación Posible:** Puntaje máximo que se podía alcanzar en la prueba.
- **Tiempo por Caso:** Número de segundos que el médico tardó en responder cada caso clínico.
- **Manejo (%)**: Porcentaje de aciertos en preguntas sobre decisiones de manejo clínico.
- **Conocimiento Factual (%)**: Porcentaje de aciertos en preguntas de conocimiento médico puro.
- **Diagnóstico (%)**: Porcentaje de respuestas correctas en preguntas sobre diagnóstico.
- **Específico (%)**: Precisión en preguntas que requerían respuestas detalladas sobre el caso.
- **General (%)**: Precisión en preguntas de conocimiento médico general.