Rev Clin Esp. 2023;224(Supl 2):S1072

Revista Clínica Española

https://www.revclinesp.es

2161 - INTERNISTAS Y ESTUDIANTES FRENTE A LA INTELIGENCIA ARTIFICIAL: COMPARATIVA DE RESULTADOS SOBRE UN CUESTIONARIO CLÍNICO

Laura Ibarra Veganzones, *Anatolio Crespo Alonso, Antonio Terrón Muñiz, Marta Obra Pinacho, Carlos Heredia Mena y Eduardo Miguel Aparicio Minguijón*

Hospital Universitario 12 de Octubre, Madrid, España.

Resumen

Objetivos: La aparición de sistemas de lenguaje basados en el aprendizaje supervisado y mediante refuerzo, comúnmente conocidos como "Inteligencia artificial" (IA), pone en cuestión la vigencia del aprendizaje, evaluación y ejercicio de la medicina. El objetivo del estudio es definir la capacidad discriminativa de dos de estos sistemas de IA (Chat GPT 3.5 y Bing Chat) para responder adecuadamente preguntas médicas.

Métodos: Se ha sometido a un grupo de 17 personas (estudiantes de Medicina, residentes de Medicina Interna y Facultativos Especialistas en Medicina Interna) del Hospital Universitario 12 de Octubre y Facultad de Medicina de la Universidad Complutense de Madrid a un cuestionario de 24 preguntas sobre conocimiento médico general. Las preguntas se han clasificado según su dificultad (fácil/intermedia/difícil) y el tipo de formulación (teórica/caso clínico). Las respuestas se han clasificado en según su precisión de acierto en: correcta, incorrecta y respuesta próxima a la correcta. Se ha comparado el grado de acierto de las respuestas elaboradas de manera individual por el grupo previamente mencionado con la aportada por Chat GPT 3.5 y Bing Chat. Para ello se ha realizado un análisis estadístico mediante U de Mann Whitney para variables independientes por parte de un investigador que desconocía las entidades de cada respuesta.

Resultados: La puntuación media obtenida en el grupo de personas encuestadas fue de 56,9 puntos (sobre un total de 72), frente a la puntuación media de 59 puntos obtenida por el grupo de las IA (p 0,83). La puntuación media en las preguntas teóricas fue de 37,8 puntos (sobre un total de 48) para el grupo de personas encuestadas frente a la puntuación media de 42 obtenida por el grupo de las IA, sin alcanzar la significación estadística. Respecto a los casos clínicos, las personas encuestadas obtuvieron una media de 19,1 puntos (sobre 24) mientras que las IA obtuvieron una media de 17 puntos, sin alcanzar la significación estadística.

Discusión: Según el análisis realizado, no se encontraron diferencias en la tasa de acierto de ninguno de los dos programas de Inteligencia Artificial respecto al grupo general de encuestados, si bien sí se observó una tendencia numérica favoreciendo a las IA en las preguntas teóricas y los humanos en los casos clínicos, sin alcanzar la significación estadística. En cualquier caso, no podemos ignorar que el pequeño tamaño muestral obtenido limita la validez de dichos resultados. En cuanto a los objetivos secundarios, no se encontraron diferencias entre la tasa de acierto de ambos

programas. Además, cabe destacar que la categorización de acierto fue determinada por la opinión subjetiva de los investigadores. Como fortaleza, nuestro estudio refuerza la teórica capacidad de los sistemas de lenguaje de IA para resolver problemas clínicos habituales. No obstante, se necesitan más estudios con muestras del tamaño suficiente para la comprobación de dicha teoría.

Conclusiones: Nuestro estudio no encontró diferencias significativas entre la tasa de éxito de respuestas a preguntas médicas clínicas entre los programas de inteligencia artificial y el personal médico/estudiantes de Medicina. Si bien, es necesario realizar más estudios con tamaño muestral suficiente para sacar conclusiones robustas.