



Precisión de la inteligencia artificial como método de apoyo diagnóstico en la consulta de otorrinolaringología

Accuracy of artificial intelligence as a diagnostic support tool in the otorhinolaryngology outpatient consultation.

Correspondencia

Raúl F Delgado Nieves
rauldelgado@lasallistas.org.mx

ORCID

<https://orcid.org/0009-0008-8273-3959>
<https://orcid.org/0009-0004-8803-9977>

Recibido: 26 de junio 2026

Aceptado: 9 de septiembre 2026

Este artículo debe citarse como: Delgado-Nieves RF, Pineda-Cásarez F, Durán-Ortiz M. Precisión de la inteligencia artificial como método de apoyo diagnóstico en la consulta de otorrinolaringología. An Orl Mex 2026; 71 (4): 299-307. <https://doi.org/10.24245/aorl.v71i4.10988>

PARA DESCARGA

<https://doi.org/10.24245/aorl.v71i4.10988>

<https://otorrino.org.mx>
<https://nietoeditores.com.mx>

Raúl F Delgado Nieves, Fernando Pineda Cásarez, Mariana Durán Ortiz

Departamento de Otorrinolaringología, Hospital Regional General Ignacio Zaragoza, ISSSTE. Facultad Mexicana de Medicina, Universidad La Salle, Ciudad de México.

Resumen

OBJETIVOS: Evaluar la concordancia diagnóstica entre una herramienta de inteligencia artificial (IA) basada en modelo de lenguaje extenso y el juicio clínico de médicos otorrinolaringólogos en pacientes con enfermedades frecuentes de esta especialidad, así como analizar su efecto en la confianza diagnóstica del médico y la percepción de utilidad clínica.

MATERIALES Y MÉTODOS: Estudio observacional, transversal y descriptivo, en el que se analizaron expedientes clínicos electrónicos. Se incluyeron médicos otorrinolaringólogos y residentes que evaluaron casos clínicos reales de rinitis, otitis media aguda, vértigo postural paroxístico benigno y disfonía. Se comparó el diagnóstico clínico emitido por el médico con el diagnóstico sugerido por una herramienta de inteligencia artificial (ChatGPT-4). Se midió la concordancia diagnóstica, la confianza del médico y la percepción de utilidad mediante cuestionarios estructurados.

RESULTADOS: La concordancia diagnóstica global entre la IA y el médico especialista fue del 68% de coincidencia bruta; sin embargo, el índice kappa de Cohen fue de 0.132 (concordancia escasa) para el primer paso diagnóstico y de 0.235 (concordancia discreta) tras la incorporación de estudios de extensión. La confianza diagnóstica pre y pos-IA no tuvo diferencia estadísticamente significativa.

CONCLUSIONES: La herramienta de IA evaluada mostró una concordancia diagnóstica escasa a discreta con el juicio clínico del especialista. A pesar de su limitada

concordancia, los médicos en formación la percibieron como útil. La IA no debe considerarse sustituto del juicio clínico especializado, sino un complemento.

PALABRAS CLAVE: Aprendizaje automático; lenguaje; decisiones clínicas; informática médica; inteligencia artificial.

Abstract

OBJECTIVES: To evaluate the diagnostic agreement between an artificial intelligence (AI) tool based on a large language model and the clinical judgment of otorhinolaryngologists in patients with common conditions in this specialty, as well as to analyze its impact on physicians' diagnostic confidence and perceived clinical utility.

MATERIALS AND METHODS: Observational, cross-sectional, descriptive study, in which 100 anonymized electronic clinical records were analyzed. Otolaryngologists and residents were included and evaluated real clinical cases of rhinitis, acute otitis media, benign paroxysmal positional vertigo, and dysphonia. The clinical diagnosis issued by the physician was compared with the diagnosis suggested by an artificial intelligence tool (ChatGPT-4). Diagnostic agreement, physician confidence, and perceived usefulness were measured using structured questionnaires.

RESULTS: The global diagnostic agreement between AI and the specialist physician was 68% raw concordance; however, the Cohen's kappa index was 0.132 (poor agreement) for the first diagnostic step and 0.235 (slight agreement) after incorporating extension studies. Pre-AI and post-IA diagnostic confidence had no statistically significant difference.

CONCLUSIONS: The evaluated artificial intelligence tool showed poor to slight diagnostic agreement with the specialist's clinical judgment in otorhinolaryngology. Despite its limited agreement, it was perceived as useful by physicians in training. AI should not be considered a substitute for specialized clinical judgment, but rather as a complement.

KEYWORDS: Machine learning; Language; Clinical decision-making; Medical informatics; Artificial intelligence.

ANTECEDENTES

En los últimos años, los modelos de inteligencia artificial (IA), particularmente los modelos de lenguaje extenso (*large language models*), han demostrado capacidades significativas en el procesamiento de información clínica y el razonamiento diagnóstico asistido.^{1,2} El surgimiento de plataformas como ChatGPT (OpenAI, San Francisco, California) ha generado un creciente interés en su aplicación potencial en el ámbito de la salud, incluidos escenarios de diagnóstico diferencial, educación médica y apoyo a la toma de decisiones clínicas.^{3,4}

No obstante, la validación clínica rigurosa de estas herramientas permanece limitada. La mayor parte de los estudios disponibles se han centrado en su aplicación para fines educativos, como generación de resúmenes y guías de estudio, mientras que la evidencia respecto de su desempeño frente a casos clínicos reales en entornos de consulta externa sigue siendo escasa.⁵ En el contexto latinoamericano, y específicamente en México, la investigación al respecto es particularmente limitada, lo que constituye una brecha significativa en la bibliografía médica regional.

El concepto de concordancia diagnóstica, entendido como el grado de acuerdo entre dos o más métodos de clasificación o evaluación, difiere sustancialmente del de precisión diagnóstica. La precisión implica la comparación contra un estándar de referencia validado (criterio de verdad), mientras que la concordancia evalúa el grado de acuerdo entre observadores o métodos, sin implicar necesariamente la corrección del diagnóstico. En este estudio, debido a que el diagnóstico del médico especialista se estableció como estándar de referencia, el análisis corresponde propiamente a una evaluación de concordancia intermétodo entre la IA y el juicio clínico médico.⁶

El índice kappa de Cohen constituye el estadístico de elección para evaluar la concordancia interobservadores porque corrige el acuerdo esperado por azar. La interpretación de sus valo-

res sigue la clasificación de Landis y Koch: los valores menores de 0.20 indican concordancia escasa; 0.21-0.40, discreta; 0.41-0.60, moderada; 0.61-0.80, buena, y 0.81-1.00, muy buena.⁷ Es fundamental destacar que una coincidencia porcentual elevada (por ejemplo, 68%) no necesariamente se traduce en una concordancia aceptable si el acuerdo atribuible al azar es alto, situación que el índice kappa permite identificar.

El objetivo de este estudio fue evaluar la concordancia diagnóstica entre ChatGPT-4 y el juicio clínico de médicos otorrinolaringólogos en pacientes con enfermedades frecuentes de esta especialidad (rinitis, otitis media aguda, vértigo postural paroxístico benigno y disfonía), así como analizar el efecto de la exposición a la IA en la confianza diagnóstica de médicos residentes y su percepción de utilidad clínica.

MATERIALES Y MÉTODOS

Estudio observacional, analítico y transversal, conforme a las recomendaciones de la guía STROBE (*Strengthening the Reporting of Observational Studies in Epidemiology*).⁸ El estudio lo aprobó el Comité de Ética de la institución, con autorización para el análisis de datos clínicos obtenidos de forma rutinaria y anonimizados, por lo que no fue necesario el consentimiento informado adicional. Los procedimientos siguieron las normas éticas establecidas por la Asociación Médica Mundial en la Declaración de Helsinki.

Población y muestra

La población de estudio estuvo conformada por médicos residentes de tercer (R3) y cuarto (R4) año de la especialidad de otorrinolaringología que aceptaron participar voluntariamente. Se recopiló una muestra consecutiva de 100 expedientes clínicos electrónicos de pacientes atendidos en seguimiento en el servicio de otorrinolaringología del Hospital Regional General Ignacio Zaragoza (ISSSTE), institución de tercer nivel en la Ciudad de México, durante el periodo del 1 de enero al 30 de junio de 2024.

Criterios de selección

Criterios de inclusión: 1) pacientes mayores de 18 años cuyo expediente contara con motivo de consulta, antecedentes y exploración física completos. 2) Pacientes con seguimiento documentado de al menos 6 meses. 3) casos clínicos correspondientes a las siguientes enfermedades: rinitis alérgica, otitis media aguda, vértigo posicional paroxístico benigno y disfonía funcional.

Criterios de exclusión: 1) expedientes con datos incompletos. 2) pacientes con diagnóstico previo conocido que pudiera condicionar el razonamiento diagnóstico. 3) Casos con enfermedades concomitantes que impidieran establecer un diagnóstico principal único.

Procedimiento

Se estandarizó un protocolo de introducción de datos clínicos en el modelo ChatGPT-4 (OpenAI, San Francisco, California, versión 2024). En cada sesión se procesaron tres casos clínicos siguiendo una secuencia bifásica:

Primera fase (diagnóstico basado solo en datos clínicos): se introdujeron los datos del motivo de consulta, antecedentes y exploración física obtenidos de la historia clínica de primera vez mediante el siguiente *prompt* estandarizado: “Responde como médico especialista en otorrinolaringología. Con base en la siguiente presentación clínica, proporciona los tres diagnósticos diferenciales más probables, ordenados por probabilidad”. Se registró el primer diagnóstico diferencial proporcionado por la IA como diagnóstico principal.

Segunda fase (diagnóstico con estudios complementarios): posteriormente, se introdujeron los datos de los estudios de extensión (audiometría, tomografía computada, rinoscopia, etc.) cuando estaban disponibles, utilizando el siguiente *prompt*: “Actúa como médico especialista en otorrinolaringología. Con base en la manifestación clínica anterior e integrando los siguientes estudios complementarios, proporciona el diagnóstico más probable”.

El diagnóstico otorgado por el médico tratante en la historia clínica se estableció como estándar de referencia para fines de comparación. La concordancia se definió como la coincidencia exacta entre el diagnóstico principal emitido por la IA y el diagnóstico final establecido por el médico tratante.

Evaluación por médicos residentes

De manera paralela, los casos clínicos se presentaron a los médicos residentes participantes (R3, n = 4; R4, n = 3), quienes registraron: *a*) su diagnóstico independiente; *b*) confianza diagnóstica pre-IA en escala Likert de 1 a 5 (1: muy baja, 5: muy alta); *c*) confianza diagnóstica pos-IA tras conocer la sugerencia de la IA; *d*) percepción de utilidad de la sugerencia de la IA en escala Likert de 1 a 5 (1: nada útil, 5: muy útil).

Variables de estudio

Variables principales: concordancia diagnóstica IA-médico (dicotómica: concordante-discordante), índice kappa de Cohen con intervalo de confianza del 95%. *Variables secundarias*: confianza diagnóstica pre y pos-IA, utilidad percibida, distribución por enfermedad y nivel de experiencia (R3 vs R4).

Análisis estadístico

Los datos se organizaron en hoja de cálculo Microsoft Excel y se analizaron con el programa IBM SPSS Statistics version 25 (IBM Corp., Armonk, NY, Estados Unidos). Se realizó estadística descriptiva para variables cualitativas (frecuencias y porcentajes) y cuantitativas (media $\pm$ desviación estándar o mediana [rango intercuartílico] según distribución).

Para evaluar la concordancia diagnóstica se calculó el índice kappa de Cohen con su intervalo de confianza del 95%, utilizando la clasificación de Landis y Koch para su interpretación.⁷ La concordancia se definió como la coincidencia exacta entre el diagnóstico principal emitido por la IA y el final establecido por el médico tratante.

Para las variables continuas relacionadas con confianza diagnóstica se utilizó la prueba t de Student para muestras pareadas (pre vs pos-IA) y la prueba U de Mann-Whitney para la comparación entre grupos independientes (R3 vs R4). Para las variables categóricas se utilizó la prueba χ^2 o exacta de Fisher según correspondiera. Un valor de $p < 0.05$ se consideró estadísticamente significativo.

El tamaño de la muestra (n = 100) se determinó considerando la estimación de concordancia con un margen de error del 10% y un nivel de confianza del 95%, asumiendo una concordancia esperada del 70%.

RESULTADOS

Se analizaron 100 expedientes clínicos que cumplieron con todos los criterios de selección. La distribución por enfermedad fue: rinitis alérgica (30%), disfonía funcional (26%), vértigo

posicional paroxístico benigno (24%) y otitis media aguda (20%). Participaron siete médicos residentes: cuatro de tercer año (R3) y tres de cuarto año (R4). Las características de la muestra se resumen en el **Cuadro 1**.

Confianza diagnóstica y utilidad percibida

La confianza diagnóstica autopercebida por los médicos residentes previa a la exposición a la IA fue de 3.92 ± 0.78 en una escala de 1 a 5. Posterior a la visualización del diagnóstico sugerido por la IA, la confianza fue de 3.80 ± 0.85 , lo que evidencia una diferencia no estadísticamente significativa (prueba t pareada, $p > 0.05$). La magnitud de esta diferencia (-0.12 puntos) carece de relevancia clínica significativa.

Al analizar por nivel de experiencia, no se encontraron diferencias significativas en los niveles de confianza pre-IA ni pos-IA entre residentes de tercer y cuarto año (prueba U de Mann-Whitney, $p > 0.05$). La utilidad percibida del apoyo de la IA fue de 3.93 ± 0.82 , lo que se considera moderada-alta. No se identificaron diferencias significativas en la utilidad percibida entre los distintos tipos de enfermedad (prueba de Kruskal-Wallis, $p = 0.155$). Los resultados se muestran en el **Cuadro 2**.

Concordancia diagnóstica

Las coincidencias diagnósticas entre la IA y el médico especialista en la primera fase (evaluación basada únicamente en datos clínicos) fueron del 68% de concordancia bruta. Sin embargo, el análisis mediante el índice kappa de Cohen reveló una concordancia escasa (kappa = 0.132; IC95%: -0.067 a 0.331). Tras la incorporación de los estudios de extensión (segunda fase), la coincidencia bruta aumentó al 79%, con un índice kappa de 0.235 (IC95%: 0.048 a 0.422),

Cuadro 1. Características de la muestra y estadísticas descriptivas (n = 100)

Variable	Total, n = 100 (%)	Vértigo posicional paroxístico benigno, n = 24 (%)	Disfonía, n = 26 (%)	Otitis media aguda, n = 20 (%)	Rinitis, n = 30 (%)
Edad (años)	45.2 ± 18.3	52.1 ± 16.8	42.3 ± 19.1	38.7 ± 17.5	47.6 ± 18.9
Sexo femenino	58 (58)	16 (64)	14 (52)	10 (53)	18 (62)
Confianza pre-IA*	3.92 ± 0.78	3.96 ± 0.81	3.89 ± 0.75	3.95 ± 0.79	3.88 ± 0.80
Confianza pos-IA*	3.80 ± 0.85	3.85 ± 0.82	3.79 ± 0.78	3.81 ± 0.86	3.73 ± 0.87
Utilidad percibida*	3.93 ± 0.82	3.84 ± 0.80	3.93 ± 0.83	4.26 ± 0.73	3.79 ± 0.84

Los datos se muestran como media ± desviación estándar o frecuencia (porcentaje).

* Escala Likert 1-5.

Cuadro 2. Confianza diagnóstica y utilidad percibida de la IA (n = 100)

Variable	Media ± DE
Confianza diagnóstica pre-IA	3.92 ± 0.78
Confianza diagnóstica pos-IA	3.80 ± 0.85
Utilidad percibida de la IA	3.93 ± 0.82

Los datos se presentan como media ± desviación estándar en escala Likert 1-5.

correspondiente a una concordancia discreta según la clasificación de Landis y Koch.⁷ En ambas fases los límites inferiores de los intervalos de confianza incluyen el valor 0, lo que indica que la concordancia observada no es estadísticamente diferente del acuerdo por azar.

Estos hallazgos demuestran que una coincidencia porcentual aparentemente moderada (68%) puede enmascarar una concordancia realmente baja una vez corregido el efecto del azar, lo que subraya la importancia de reportar el índice kappa junto con los porcentajes de coincidencia bruta en estudios de concordancia diagnóstica. **Cuadro 3**

Concordancia por enfermedad

El análisis de concordancia estratificado por enfermedad (enfoque *one-vs-rest*) evidenció variabilidad en los resultados. Los valores de kappa más elevados se observaron en disfonía (kappa = 0.41; IC95%: 0.18-0.64; concordancia moderada) y vértigo posicional paroxístico benigno (kappa = 0.46; IC95%: 0.22-0.70; concordancia moderada). Por el contrario, la otitis media aguda mostró el menor valor de concordancia (kappa = 0.33; IC95%: 0.08-0.58; concordancia discreta), seguida de la rinitis (kappa = 0.29; IC95%: 0.05-0.53; concordancia discreta). No obstante, estas diferencias no alcanzaron significación estadística al compararse mediante la prueba χ^2 ($p > 0.05$), lo que sugiere que el desempeño de la IA fue relativamente homogéneo entre las distintas enfermedades evaluadas. Los resultados se detallan en el **Cuadro 4**.

DISCUSIÓN

Este estudio evaluó la concordancia diagnóstica entre una herramienta de inteligencia artificial basada en modelo de lenguaje extenso (ChatGPT-4) y el juicio clínico de médicos especialistas en otorrinolaringología, así como el efecto de la exposición a la IA en la confianza diagnóstica de médicos residentes.

Los resultados principales indican que la concordancia diagnóstica entre la IA y los médicos especialistas fue escasa a discreta (kappa = 0.132-0.235), valores considerablemente inferiores a los reportados en estudios de otras especialidades médicas. Por ejemplo, en dermatología, Esteva y su grupo⁹ reportaron valores de kappa entre 0.60 y 0.75 para sistemas de clasificación

Cuadro 3. Concordancia entre IA y médico especialista: índice kappa de Cohen

Comparación	Coincidencia (n)	Coincidencia (%)	Índice kappa	IC95%	Interpretación
IA vs médico especialista, datos basales	68	68	0.132	-0.067 a 0.331	Escasa
IA vs médico especialista, estudios de extensión	79	79	0.235	0.048 a 0.422	Discreta

Cuadro 4. Concordancia diagnóstica por enfermedad

Enfermedad	N	k	IC95%	Interpretación
Rinitis	30	0.29	0.05-0.53	Discreta
Disfonía	26	0.41	0.18-0.64	Moderada
Vértigo posicional paroxístico benigno	24	0.46	0.22-0.70	Moderada
Otitis media aguda	20	0.33	0.08-0.58	Discreta

de lesiones cutáneas mediante redes neuronales profundas. De manera similar, en radiología, McKinney y colaboradores¹⁰ encontraron concordancias moderadas a buenas ($\kappa = 0.50-0.70$) en sistemas de detección de cáncer de mama. Las diferencias observadas pueden atribuirse a múltiples factores: los sistemas de IA evaluados en dermatología y radiología se entrenan con bases de datos especializadas de imágenes etiquetadas por múltiples expertos, mientras que los modelos de lenguaje extenso generales procesan información textual narrativa sin adiestramiento específico en patrones otorrinolaringológicos.¹¹

Un hallazgo particularmente relevante es la discrepancia entre la coincidencia porcentual bruta (68-79%) y el índice kappa corregido por azar (0.132-0.235). Esta divergencia ilustra por qué el reporte exclusivo de porcentajes de acuerdo puede conducir a interpretaciones erróneas en estudios de concordancia. La coincidencia bruta no considera el acuerdo esperado por azar que, en este caso, fue elevado debido a la prevalencia desigual de las enfermedades incluidas. Este fenómeno, conocido como paradoja de Kappa, es bien documentado en la bibliografía metodológica y subraya la necesidad de reportar siempre el índice kappa con su intervalo de confianza.^{6,7}

Respecto del efecto en la confianza diagnóstica de los médicos residentes, se observó una disminución no significativa de 0.12 puntos en la escala Likert tras la exposición a la sugerencia de la IA (3.92 *vs* 3.80; $p > 0.05$). Este hallazgo es metodológicamente relevante por dos razones: primero, indica que los médicos en formación no aceptan de manera acrítica las sugerencias generadas por la IA, manteniendo su autonomía clínica y pensamiento crítico; segundo, sugiere que la discordancia entre el diagnóstico propio y el de la IA genera un proceso de reevaluación diagnóstica que, lejos de ser perjudicial, puede fomentar el razonamiento clínico reflexivo.¹² Este comportamiento concuerda con investigaciones previas acerca de la adopción de tecnologías de soporte a decisiones clínicas, donde los médicos experimentados tienden a comparar las recomendaciones automatizadas con su propio juicio clínico antes de modificar sus decisiones.¹³

La utilidad percibida de la IA, calificada como moderada-alta (3.93/5), sugiere que los médicos en formación reconocen el valor potencial de estas herramientas como complemento al razonamiento clínico, aun cuando su concordancia diagnóstica sea limitada. Esta percepción positiva podría relacionarse con la utilidad de la IA para generar diagnósticos diferenciales amplios o para acceder rápidamente a información actualizada, funciones distintas a la emisión de un diagnóstico definitivo.

Limitaciones

Este estudio tiene limitaciones que deben considerarse al interpretar los resultados. Primera, el tamaño de la muestra ($n = 100$), aunque adecuado para una estimación inicial, limita la generalización de los resultados y el poder estadístico para detectar diferencias en subgrupos. Segunda, la ausencia de un estándar de referencia independiente para los diagnósticos impide determinar si las discordancias observadas se deben a errores de la IA, del médico especialista o de ambos. El diagnóstico del médico tratante, aunque constituye el criterio práctico habitual, puede incorporar sesgos propios del adiestramiento institucional y de la experiencia individual. Tercera, el diseño unicéntrico puede introducir sesgos relacionados con los protocolos diagnósticos específicos de la institución y las características de la población atendida. Cuarta, el uso de un modelo de IA general (ChatGPT-4) en lugar de un sistema entrenado específicamente con datos otorrinolaringológicos puede subestimar el potencial real de la IA en esta especialidad.

Implicaciones para la práctica clínica e investigación futura

Los resultados sugieren que, en su estado actual de desarrollo, los modelos de lenguaje extenso de propósito general no deben utilizarse como herramientas de diagnóstico independiente en otorrinolaringología. Su utilidad potencial radica más bien en funciones de soporte: generación de diagnósticos diferenciales, acceso a información actualizada y estimulación del razonamiento crítico. Las futuras investigaciones deberán dirigirse al desarrollo y validación de modelos de IA adiestrados específicamente con datos otorrinolaringológicos, con diseños multicéntricos que incluyan estándares de referencia diagnósticos independientes (confirmación histopatológica, imagenología de referencia, seguimiento prolongado) y ensayos clínicos aleatorizados que evalúen el efecto de la IA en resultados clínicos relevantes.

CONCLUSIONES

Este estudio proporciona evidencia inicial de la integración de herramientas de inteligencia artificial basadas en modelos de lenguaje extenso en el proceso diagnóstico otorrinolaringológico. La herramienta de inteligencia artificial evaluada (ChatGPT-4) mostró una concordancia diagnóstica escasa a discreta con el juicio clínico del especialista en otorrinolaringología ($\kappa = 0.132-0.235$), lo que indica que su desempeño diagnóstico actual no es suficiente para su uso independiente en la práctica clínica. La coincidencia porcentual bruta (68-79%) no reflejó adecuadamente la magnitud real de la concordancia, lo que evidencia una importante discrepancia con los valores del índice kappa corregido por azar. Este hallazgo subraya la necesidad metodológica de reportar el índice kappa como medida primaria de concordancia en estudios similares. El uso de la IA no afectó significativamente la confianza diagnóstica de los médicos en formación, lo que sugiere que éstos mantienen su autonomía clínica y criterio crítico frente a las sugerencias automatizadas.

La percepción de utilidad fue moderada-alta (3.93/5), lo que indica un reconocimiento del valor potencial de la IA como complemento del razonamiento clínico, aunque no como sustituto del juicio especializado. El desempeño de la IA fue relativamente homogéneo entre las diferentes enfermedades evaluadas, sin diferencias significativas por tipo de afección ni por nivel de experiencia del médico evaluador. En consecuencia, se recomienda que las herramientas de IA en la práctica otorrinolaringológica se implementen con precaución, reconociendo sus limitaciones actuales y priorizando su uso como apoyo al razonamiento clínico bajo supervisión especializada, en lugar de como instrumento diagnóstico autónomo.

DECLARACIONES

Conflicto de interés

No existe ningún tipo de conflicto de interés directo o indirecto relacionado con el tema del trabajo.

Financiamiento

No se recibió ningún patrocinio para llevar a cabo la investigación reportada en este manuscrito. La investigación se realizó con recursos personales usando las instalaciones institucionales del Hospital Regional General Ignacio Zaragoza, ISSSTE.

Agradecimientos

A todo el personal del servicio de Otorrinolaringología, en especial a los médicos residentes, del Hospital Regional General Ignacio Zaragoza, ISSSTE, por su apoyo en la identificación, participación y recopilación de expedientes clínicos.

REFERENCIAS

1. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023; 2 (2): e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
2. Singhal K, Tu T, Gottweis J, et al. Large language models encode clinical knowledge. *Nature* 2023; 620 (7972): 172-180. <https://doi.org/10.1038/s41586-023-06291-2>
3. Jiang F, Liu Y, Jiang Y. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017; 2 (4): 230-243. <https://doi.org/10.1136/svn-2017-000101>
4. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J* 2019; 6 (2): 94-98. <https://doi.org/10.7861/futurehosp.6-2-94>
5. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the Chat-GPT model. *Res Sq* 2023; rs.3.rs-2566942. <https://doi.org/10.21203/rs.3.rs-2566942/v1>
6. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005; 85 (3): 257-268.
7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33 (1): 159-174. <https://doi.org/10.2307/2529310>
8. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *Ann Intern Med* 2007; 147 (8): 573-577. <https://doi.org/10.7326/0003-4819-147-8-200710160-00010>
9. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542 (7639): 115-118. <https://doi.org/10.1038/nature21056>
10. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020; 577 (7788): 89-94. <https://doi.org/10.1038/s41586-019-1799-6>
11. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; 25 (1): 44-56. <https://doi.org/10.1038/s41591-018-0300-7>
12. Clancey WJ. The epistemology of a rule-based expert system: a framework for explanation. *Artif Intell* 1983; 20 (3): 215-251. [https://doi.org/10.1016/0004-3702\(83\)90008-5](https://doi.org/10.1016/0004-3702(83)90008-5)
13. Sutton RT, Pincock D, Baumgart DC, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020; 3: 17. <https://doi.org/10.1038/s41746-020-0221-y>
14. OpenAI. GPT-4 technical report. *arXiv* 2023; 2303.08774.
15. Sarraju A, Bruemmer D, Van Iterson E, et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* 2023; 329 (10): 842-844. <https://doi.org/10.1001/jama.2023.1044>