REUE | Original

Evaluación de los conocimientos de varios sistemas de inteligencia artificial sobre una subespecialidad de la medicina de urgencias y emergencias: la toxicología clínica

Santiago Nogué-Xarau[1], Montserrat Amigó-Tadín[2], José Ríos-Guillermo[3]

OBJETIVO. La inteligencia artificial (IA) es una disciplina de la informática que se encarga de crear sistemas capaces de realizar tareas que se atribuyen a la inteligencia humana. El objetivo principal de este estudio ha sido evaluar las respuestas de algunas IA a preguntas del campo de la toxicología clínica (TC).

MATERIAL Y MÉTODOS. Se han valorado cuatro aplicaciones de IA: ChatGPT, Bing, LuzIA y Bard. Para evaluar sus conocimientos en TC se les formularon 30 preguntas sobre diversos aspectos de la TC. Cada pregunta ofrecía cinco opciones de respuesta, de las cuales sólo una era correcta. Se evaluó el acierto/error en la respuesta, así como si había apoyo bibliográfico. Si se detectaban respuestas erróneas, se reformuló la misma pregunta, pero utilizando otra forma de lenguaje para evaluar de nuevo la respuesta y ver si la misma era sensible a la calidad de la pregunta. Los datos se introdujeron en una base SPSS para su análisis estadístico. Se consideró significativo un valor de $p < 0,05$.

RESULTADOS. Los porcentajes de respuestas acertadas fueron del 70% (Bing), 67% (ChatGPT y LuzIA) y 57% (Bard), sin diferencias estadísticamente significativas. Al reformular las preguntas en los casos en los que la respuesta de la IA había sido errónea, los porcentajes de aciertos subieron en los cuatro sistemas, pero sin diferencias significativas. En sus respuestas, Bing ofreció el acceso directo a tres citas bibliográficas y Bard a cuatro, pero su presencia en PubMed era muy baja (7,2% y 0,85% respectivamente).

CONCLUSIONES. Los cuatro sistemas de IA han mostrado una capacidad de acierto en más del 50% de las preguntas formuladas de TC. No obstante, el soporte bibliográfico que proporcionan es escaso y de muy baja calidad.

Palabras clave: Inteligencia artificial. Toxicología. Docencia. Preguntas de examen.

Assessing 4 artificial intelligence systems' knowledge of a subspecialty of emergency medicine: clinical toxicology

BACKGROUND AND OBJECTIVE. Artificial intelligence (AI) is a branch of computer technology that develops systems able to perform tasks associated with human intelligence. The main objective of this study was to evaluate AI answers to questions related to clinical toxicology.

MATERIALS AND METHODS. We evaluated 4 AI applications: ChatGPT, Bing, LuzIA, and Bard. Thirty multiple-choice test questions in Spanish about various aspects of clinical toxicology were presented to the applications, and the answers were assessed. Each question included 5 possible answers, 1 of which was correct. In addition to correctness, we evaluated the bibliographic support each application provided. If the application gave an incorrect answer, we rephrased the question, presented it again, and reevaluated the new answer to detect whether question quality influenced performance. Data were recorded for analysis with SPSS. The level of statistical significance was set at $P < .05$.

RESULTS. The scores achieved by the AI applications were as follows: Bing, 70%; ChatGPT and LuzIA, 67% each; and Bard, 57% ($P > .05$). The scores improved after the incorrect questions were rephrased, but the differences were not significant. Bing included direct access to 3 references per question and Bard to 4. However, only 7.2% and 0.85% of the references, respectively, were to PubMed-indexed sources.

CONCLUSIONS. All 4 AI applications were able to correctly answer more than half the questions about clinical toxicology. After rephrasing some questions, each system achieved more correct answers. The supporting references the applications provided were few and of poor quality.

Keywords: Artificial intelligence. Toxicology. Education. Test questions.

Filiación de los autores: [1]Fundación Española de Toxicología Clínica, Barcelona, España. [2]Área de Urgencias, Hospital Clínic, Barcelona, España. [3]Departamento de Farmacología Clínica, Hospital Clínic, Barcelona, España.
Correspondencia: Santiago Nogué-Xarau. Fundación Española de Toxicología Clínica.
E-mail: snoguex@gmail.com
Información del artículo: Recibido: 21-12-2023. Aceptado: 26-12-2023. *Online:* 3-1-2024.
Editor responsable: Guillermo Burillo-Putze.

Introducción

La inteligencia artificial (IA) es una nueva disciplina de la informática que se caracteriza por su capacidad para realizar tareas que simulan a las que realiza la inteligencia humana y entre las que se encuentran la rápida resolución de problemas complejos[1]. Estos sistemas necesitan una enorme cantidad de datos para su entrenamiento, así como algoritmos y modelos matemático-estadísticos que les permitan procesar esta información, y todo ello alojado en supercomputadores dotados de una gran capacidad de almacenamiento y tratamiento de datos, de forma que pueden realizar billones de operaciones en un segundo[2].

Los usos de los sistemas de IA en la vida diaria son múltiples e incluyen aplicaciones de mapas en la web que indican la mejor ruta entre diferentes ubicaciones, asistentes virtuales *(chatbots),* filtros de *spam* en el correo electrónico, reconocimiento facial en los teléfonos móviles, sistemas de traducción automática en lenguaje natural, etc.[3].

En la docencia universitaria, la IA podría llegar a ofrecer asesoramiento personalizado a los estudiantes mediante *chatbots* que pueden responder a sus dudas, orientarlos sobre los recursos disponibles o sugerirles estrategias de aprendizaje[4]. La IA también puede ayudar a la calificación de los exámenes mediante sistemas que permiten detectar más fácilmente el plagio o mostrar errores que se podrían pasar por alto en la corrección por un humano al no estar sujeta a factores como el cansancio.

En el campo de las ciencias de la salud, la IA podría ayudar a identificar factores de riesgo de padecer enfermedades en una determinada población al realizar el análisis de sus datos sanitarios. Podría también ayudar a revisar pruebas de imagen como mamografías para complementar al radiólogo en el despistaje de imágenes sospechosas de cáncer, y podría mejorar la asistencia a los pacientes mediante dispositivos electrónicos que pueden detectar arritmias cardiacas[5-6], entre otras utilidades. En la medicina de urgencias y emergencias (MUE), se han evaluado herramientas de ayuda para sugerir un diagnóstico entre las enfermedades más prevalentes en un servicio de urgencias, con resultados prometedores, pero que necesitan supervisión por parte del equipo clínico-asistencial[7].

Una de las subespecialidades de la MUE es la toxicología clínica (TC)[8], rama centrada en el diagnóstico y tratamiento de las intoxicaciones. Los sistemas de IA más populares no han sido entrenados específicamente en TC, por lo que nos fijamos, como objetivo principal de esta investigación, evaluar la capacidad de respuesta de algunas IA sobre TC y, como objetivos secundarios, averiguar si estos sistemas son sensibles a la calidad de la pregunta que se les formula y conocer la velocidad y extensión de sus respuestas, así como el potencial soporte bibliográfico que ofrecen en sus respuestas.

Material y métodos

Se han evaluado cuatro sistemas de IA que cumplían tres condiciones: acceso gratuito; intercambio instantáneo de pregunta/respuesta a través de un chat y utilización de la lengua española. Entre las diversas opciones se seleccionaron ChatGPT, Bing, LuzIA y Bard, por ser las aplicaciones con las que los autores estaban más familiarizados (Tabla 1). Todas las evaluaciones se realizaron durante el mes de mayo de 2023.

Para valorar la idoneidad de las respuestas con relación a la TC se las sometió a un examen de tipo test con 30 preguntas, que fue el mismo que tuvieron los alumnos de Toxicología de la Facultad de Medicina de la Universidad de Barcelona en la convocatoria de junio de 2019. Cada pregunta ofrecía cinco opciones de respuesta, de las cuales sólo una era correcta. Tanto las preguntas como las posibles respuestas fueron preparadas por uno de los autores (SN). Cada pregunta se incorporaba al chat de la IA y se solicitaba la respuesta que consideraba correcta entre las cinco opciones. Las respuestas de las IA, que iban a ser en formato de texto, fueron catalogadas como acierto/error por consenso entre dos de los autores expertos en TC (SN y MA).

Se evaluó el tiempo de demora en iniciar la respuesta (medido en segundos) y la cantidad de texto utilizada en la respuesta (medido en número de palabras). Se valoró también si estas respuestas iban acompañadas de soporte bibliográfico, considerando también si las citas eran falsas o verdaderas y, en este último caso, si estaban indexadas en *PubMed*.

Si se detectaban respuestas erróneas, se analizó el redactado de la pregunta y se reformuló, intentando mejorar su redacción y que no ofreciese dudas de interpretación. El nuevo enunciado, con las mismas opciones de respuesta, se le ofreció a la IA que había fallado para constatar si con ello modifica su elección.

Para el análisis estadístico (realizado por otro de los autores, JR), se introdujeron los datos en el programa SPSS (versión 26). Los resultados obtenidos se expresan en números absolutos, porcentaje (%) o mediana con su rango intercuartílico (percentiles 25 y 75) para las variables cualitativas o cuantitativas respectivamente. Las variables de acierto/error de la respuesta se evaluaron con la prueba

Tabla 1. Aspectos metodológicos del estudio: inteligencias artificiales evaluadas

Nombre	Versión	Empresa que la ha desarrollado	Vía de acceso electrónico
ChatGPT	3.5	OpenAI	https://chat.openai.com Último acceso 23 diciembre 2023
Bing	Versión disponible en mayo de 2023	Microsoft	https://www.bing.com Último acceso 23 diciembre 2023
LuzIA	Versión disponible en mayo de 2023	Fundación LuzIA	https://web.whatsapp.com Último acceso 23 diciembre 2023
Bard	1.0	Google	https://bard.google.com/chat?hl=es Último acceso 23 diciembre 2023

Tabla 2. Resultados obtenidos en las respuestas a las preguntas (n = 30) de Toxicología Clínica por las inteligencias artificiales (n = 4)

			ChatGPT	Bing	Luzia	Bard
Acierto-error al responder	Acierto	N (%)	17 (56,67)	21 (70,00)	20 (66,67)	20 (66,67)
	Error	N (%)	13 (43,33)	9 (30,00)	10 (33,33)	10 (33,33)
		Valor de p en comparación con ChatGPT		1,000	1,000	0,581
		Valor de p en comparación con Bing			1,000	0,344
		Valor de p en comparación con LuzIA				0,607
Demora en iniciar la respuesta (segundos)		Mediana [P25th; P75th]	5,78 [4,42; 6,73]	9,81 [8,00; 10,90]	21,54 [18,11; 25,61]	12,03 [7,08; 15,53]
		Rango	3,56 a 10,17	5,43 a 21,08	13,79 a 36,13	4,58 a 19,64
		N	30	30	30	30
		Valor de p en comparación con ChatGPT		< 0,001	< 0,001	< 0,001
		Valor de p en comparación con Bing			0,248	< 0,001
		Valor de p en comparación con LuzIA				< 0,001
Número de palabras utilizadas para la respuesta		Mediana [P25th, P75th]	139 [110; 168]	21 [14; 35]	139 [95; 171]	15 [12; 30]
		Rango	42 a 213	6 a 96	11 a 199	6 a 65
		N	30	30	30	30
		Valor de p en comparación con ChatGPT		< 0,001	< 0,001	0,672
		Valor de p en comparación con Bing			0,128	< 0,001
		Valor de p en comparación con LuzIA				< 0,001
Revisión y reformulación de la pregunta que se había contestado mal: acierto-error en la nueva respuesta	Acierto	N (%)	19 (63,33)	24 (80,00)	23 (76,67)	25 (83,33)
	Error	N (%)	11 (36,67)	6 (20,00)	7 (23,33)	5 (16,67)
		Valor de p en comparación con ChatGPT		1,000	0,625	0,289
		Valor de p en comparación con Bing			1,000	0,125
		Valor de p en comparación con LuzIA				0,109
		Valor de p al comparar las respuestas de cada IA, antes y después de la reformulación de las preguntas con respuesta errónea	0,250	0,250	0,063	0,500
Número de citas bibliográficas ofrecidas en cada respuesta (n[RIC])			0	3 [2,75; 4,00]	0	4 [3,00; 4.25]
Número total de citas en las 30 respuestas (n)			0	97	0	118
Citas que se comprueba que existen (n [%])			No procede	97 (100)	No procede	12 (10,17)
Citas reales procedentes de revistas médicas, farmacéuticas o de enfermería (n [% del total de citas])			No procede	24 (24,7)	No procede	1 (0,85)
Citas reales procedentes de revistas indexadas en PubMed (% del total de citas)			No procede	7 (7,2%)	No procede	1 (0,85%)

RIC: rango intercuartílico [percentiles 25 y 75].

de McNemar. Para comparar los resultados obtenidos en el tiempo de respuesta o en el número de palabras se empleó un modelo *Generalized Estimating Equation* (GEE), considerando la variabilidad intra-IA para tener en cuenta que cada pregunta se respondía por los cuatro sistemas, previa transformación de las variables cuantitativas a rango para realizar el análisis mediante una aproximación no-paramétrica. Se consideró significativo un valor de $p < 0,05$.

Resultados

Los principales resultados se muestran en la Tabla 2. Los cuatro sistemas de IA aprobaron el examen de TC, con tasas de acierto entre el 56,7% (ChatGPT) y el 70% (Bing), pero sin que hubiese diferencias estadísticamente significativas entre ellas. Los estudiantes de medicina acertaron en este mismo examen el 72% de las preguntas. Las cuatro IA contestaron bien a 17 preguntas (56,7%), las cuatro mal a 5 preguntas (16,7%) y a las 8 preguntas restantes (26,7%) unas contestaron bien y otras mal.

Los tiempos de inicio en la respuesta fueron muy rápidos, particularmente en el caso de ChatGPT (mediana de 5,78 segundos) y con diferencia respecto a los otros sistemas ($p < 0,001$). En sus respuestas, hubo dos sistemas muy expansivos en cuanto al número de palabras utilizadas (ChatGPT y LuzIA), mientras Bing y Bard fueron muy exiguos ($p < 0,001$).

Al revisar la pregunta en los casos en los que la IA había fallado y adoptar un lenguaje más preciso que no dejase dudas respecto a la respuesta correcta y volverla a pasar, se constató que los cuatro sistemas mejoraron sus resultados.

En sus respuestas, ChatGPT y LuzIA no aportaron bibliografía para dar soporte a las mismas. Bing ofreció el acceso a una mediana de 3 citas en cada respuesta; todas ellas eran reales y en un 24,7% correspondían a revistas científicas del ámbito de la salud (medicina, enfermería o farmacia), pero sólo un 7,2% estaban indexadas en PubMed. Bard aportó una mediana de 4 citas por respuesta, pero el 89,8% resultaron ser falsas ya que no se correspondía el título del trabajo, ni con los autores ni con la revista, y solo el 0,85% estaban indexadas en *PubMed*.

Discusión

Las cuatro IA han mostrado un notable conocimiento de TC, consiguiendo aprobar el examen propuesto. Los

estudiantes de medicina de la Universidad de Barcelona, acertaron en este mismo examen el 72% de las preguntas. Otros autores ya habían puesto en evidencia que estos sistemas son capaces de superar la prueba para ejercer de médico en los EE.UU. (el *United States Medical Licensing Exam* –USMLE–), ya que ChatGPT-4 contestó correctamente el 90% de las preguntas . Esta competencia de las IA indica un potencial interés para satisfacer demandas profesionales en la práctica de la medicina, especialmente cuando se manejan bases de datos con miles de pacientes[10] o cuando se está ante un caso clínico complejo[11], habiendo mostrado sus potenciales capacidades en el terreno de la medicina forense[12], salud pública[13], farmacología[14], oftalmología[15], cardiología[16], oncología[17], neumología[18], neurología[19], neurocirugía[20], hepatología[21], odontología[22], y angiología[23], entre otras especialidades[24]. Pero no hay que olvidar que los sistemas de IA estudiados, dotados de un lenguaje natural en base al enorme volumen de información disponible, no son infalibles y cometen errores derivados de la ausencia de comprensión del texto que proponen, lo que podría llevar a una mala praxis[25]. En este punto conviene recordar que la Organización Mundial de la Salud ha publicado un informe sobre ética de la IA en el ámbito de la salud, donde ofrece seis principios rectores para su concepción y utilización: proteger la autonomía humana; promover beneficios humanos; garantizar transparencia; fomentar responsabilidad; asegurar equidad; e impulsar seguridad[26]; sin embargo, no está claro que estos principios puedan cumplirse, dado que los algoritmos que hay detrás de estas IA no se hacen públicos por parte de las empresas desarrolladoras y no es posible comprobar la información de la que se nutren.

El terreno de la docencia universitaria, las IA podrían ser de ayuda tanto a los docentes como al alumnado[27,28]. Con relación a la evaluación de los alumnos, nuestro estudio muestra que ajustando las preguntas se mejoran los aciertos de las IA, lo que indica que la calidad de la pregunta impacta sobre la elección de la respuesta, aspecto que ya es conocido por el personal docente. Por ello quizás sería conveniente que los profesores pasasen previamente sus exámenes por el filtro de alguna IA y analicen las preguntas que contesta erróneamente, porque es posible que la pregunta pueda ser mejor formulada y permita, en su nuevo enunciado, que no haya dudas respecto a la respuesta correcta. Aunque, por otro lado, estas preguntas bien revisadas y, a pesar de ello, erradas por la IA son probablemente las mejores para evaluar no solo la cantidad de conocimiento que tiene el alumno sino, sobre todo, su capacidad para pensar, razonar y reflexionar sobre la respuesta adecuada, que debería ser el eje vertebral de la evaluación del alumnado en cualquier disciplina. La cara negativa de la IA es que esta podría realizar los trabajos de redacción o resumen que pida un profesor a sus alumnos. Sin embargo, aunque estas IA realizan textos de gran calidad, el hecho de no entender lo que redactan de forma automática las hace susceptibles a mostrar datos erróneos, por lo que estos sistemas no son infalibles y el alumno siempre tendrá que revisar todo lo que la IA haya vertido en un texto, comprobando la veracidad de las afirmaciones y la bibliografía aportada. Esto último queda reflejado en nuestro trabajo. El soporte bibliográfico ofrecido en sus respuestas, ha sido muy dispar, desde inexistente en dos de los sistemas, hasta aparentemente adecuado en los otros dos, con unas 3 o 4 citas en cada respuesta. Pero la calidad de esta bibliografía era discreta en el caso de Bing, con solo un 7,2% de las citas indexadas en PubMed, y muy deficiente con Bard, ya que casi el 90% de sus citas eran falsas y menos del 1% estaban indexadas en PubMed. Este bajo nivel en la bibliografía ha sido también observado por Chen *et al.* (en su estudio, un 20,6% de citas eran inexistentes)[29], hecho que pudiera explicarse por la arquitectura de estos sistemas que ponen más el foco en la generación de respuestas en lenguaje natural que en el contenido y justificación de éstas[30].

Los tiempos de respuesta por parte de las IA son extraordinariamente cortos y ésta es, junto a su insensibilidad a aspectos tan humanos como el cansancio y la condición física y psíquica, una de sus características, que facilita su utilización masiva. Aunque haya diferencias temporales significativas entre ellas, no son relevantes en la práctica, ya que en el peor de los casos la respuesta se iniciaba en unos 20 segundos.

También ha habido diferencias en el carácter expansivo de las respuestas: escaso en número de palabras en algunos casos y muy generoso en otros. Este aspecto es también poco relevante, ya que estos sistemas de IA permiten que se les solicite la respuesta con un determinado número de palabras, lo que las convierte de nuevo en muy prácticas, y por ello podrían ser utilizadas para que redacten un determinado tema con una extensión precisa.

Como limitaciones del presente trabajo cabe señalar que sólo se han realizado 30 preguntas a cada una de las IA y sobre una única especialidad como es la TC, lo que no permite extrapolar los resultados a otras ramas de la MUE, ni a otras especialidades. Además, estos sistemas de IA están evolucionando a gran velocidad y los resultados que hemos obtenido serían muy probablemente mejorados con las nuevas versiones disponibles en un futuro inmediato.

Conclusiones

Los sistemas de IA testados han mostrado sus habilidades para pasar satisfactoriamente un examen de TC, por lo que se supone que tienen acceso a conocimientos veraces y capacidad para dar respuesta a múltiples preguntas con un lenguaje correcto y entendible, por lo que podrían ser útiles para mejorar o complementar diversos aspectos relacionados con la asistencia, la docencia y la investigación en ciencias de la salud, incluida la MUE, pero teniendo siempre en cuenta los valores éticos, morales y legales en su utilización. No obstante, su incapacidad para certificar la veracidad de sus redactados hace imperativo una revisión por personal experto.

Por todo ello, es probable la implementación del uso de sistemas de IA, no sólo como parte del aprendizaje dentro del curriculum académico en los planes de estudios de todas las ciencias de la salud, sino también es aspectos de gestión, asistencia e investigación.

INFORMACIÓN DEL ARTÍCULO

Conflicto de intereses: Los autores declaran no tener conflictos de interés en relación con el presente artículo.

Financiación: Los autores declaran la no existencia de financiación externa en relación con el presente artículo.

Responsabilidades éticas: Todos los autores han confirmado el mantenimiento de la confidencialidad y respeto de los derechos de los pacientes, acuerdo de publicación y cesión de derechos de los datos a la Revista Española de Urgencias y Emergencias.

Artículo no encargado por el Comité Editorial y con revisión externa por pares.

BIBLIOGRAFÍA

1. Chat GPT. Autodefinición de inteligencia artificial. (Consultado 17 Diciembre 2023). Disponible en:
2. Howard J. Artificial intelligence: Implications for the future of work. Am J Ind Med. 2019; 62:917-26.
3. Pham KT, Nabizadeh A, Selek S. Artificial intelligence and chatbots in psychiatry. Psychiatr Q. 2022;93:249-53.
4. Zawiah M, Al-Ashwal FY, Gharaibeh L, Abu Farha R, Alzoubi KH, Abu Hammour K, et al. ChatGPT and clinical training: perception, concerns, and practice of pharm-d students. J Multidiscip Healthc. 2023;16:4099-110.
5. Tsiknakis N, Trivizakis E, Vassalou EE, Papadakis GZ, Spandidos DA, Tsatsakis A, et al. Interpretable artificial intelligence framework for COVID-19 screening on chest X-rays. Exp Ther Med. 2020;20:727-35.
6. Buelga ML, Ramírez J, Alonso GL. El reloj inteligente *(smartwatch)* ante el bloqueo auriculoventricular completo: un reto por encima de sus posibilidades. Emergencias. 2023;35:478-9.
7. Moreno E, Pueyo I, Sánchez M, Martín M, Masip J. Experiencia de Mediktor®: un nuevo evaluador de síntomas basado en inteligencia artificial para pacientes atendidos en el servicio de urgencias. Emergencias. 2017;29:391-6.
8. Nogué S. Toxicólogo y urgenciólogo: una nueva variante del cangrejo ermitaño. Emergencias. 2009;21:62-4.
9. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep. 2023;13:16492.
10. González-Díaz A, Matos-Castro S, Arruabarrena-Urrestarazu N, González-Valladares E, Molina-Padilla S, Ferrer-Dufol A, et al. Evolución de las intoxicaciones agudas por productos químicos en el quinquenio 2015-2019, registradas por el Sistema Español de Toxicovigilancia (SETv). Rev Esp Urg Emerg. 2023;2:30-5.
11. Marcos DN, Cuerpo S, Coll-Vinent B. Coma en paciente *nomen nescio*. Rev Esp Urg Emerg. 2023;2:56-7.
12. Piraianu AI, Fulga A, Musat CL, Ciobotaru OR, Poalelungi DG, Stamate E, et al. Enhancing the evidence with algorithms: how artificial intelligence is transforming forensic medicine. Diagnostics (Basel). 2023;13:2992.
13. Ayers JW, Zhu Z, Poliak A, Leas EC, Dredze M, Hogarth M, et al. Evaluating artificial intelligence responses to public health questions. JAMA Netw Open. 2023;6:e2317517.
14. Kumar M, Nguyen TPN, Kaur J, Singh TG, Soni D, Singh R, et al. Opportunities and challenges in application of artificial intelligence in pharmacology. Pharmacol Rep. 2023; 75:3-18.
15. Alam M, Hallak JA. AI-automated referral for patients with visual impairment. Lancet Digit Health. 2021;3:e2-e3.
16. Augusto JB, Davies RH, Bhuva AN, Knott KD, Seraphim A, Alfarih M, et al. Diagnosis and risk stratification in hypertrophic cardiomyopathy using machine learning wall thickness measurement: a comparison with human test-retest performance. Lancet Digit Health. 2021;3:e20-e28.
17. Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. Semin Cancer Biol. 2021;72:214-25.
18. Soffer S, Morgenthau AS, Shimon O, Barash Y, Konen E, Glicksberg BS, et al. Artificial intelligence for interstitial lung disease analysis on chest computed tomography: A systematic review. Acad Radiol. 2022; Suppl 2:S226-S235.
19. Myszczynska MA, Ojamies PN, Lacoste AMB, Neil D, Saffari A, Mead R, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Rev Neurol. 2020;16:440-56.
20. Roman A, Al-Sharif L, Al Gharyani M. The Expanding Role of ChatGPT (Chat-Generative Pre-Trained Transformer) in Neurosurgery: A systematic review of literature and conceptual framework. Cureus. 2023;15:e43502.
21. Fraser K, Bruckner DM, Dordick JS. Advancing predictive hepatotoxicity at the intersection of experimental, in silico, and artificial intelligence technologies. Chem Res Toxicol. 2018; 31:412-30.
22. Rokhshad R, Ducret M, Chaurasia A, Karteva T, Radenkovic M, Roganovic J, et al. Ethical considerations on artificial intelligence in dentistry: A framework and checklist. J Dent. 2023;135:104593.
23. Sonnenschein K, Stojanovic SD, Dickel N, Fiedler J, Bauersachs J, Thum T, et al. Artificial intelligence identifies an urgent need for peripheral vascular intervention by multiplexing standard clinical parameters. Biomedicines. 2021; 9:1456.
24. Nogue-Xarau S, Amigó-Tadin M. Ríos-Guillermo J. ¿Puede la inteligencia artificial ayudar al urgenciólogo en el diagnóstico de las intoxicaciones? Emergencias 2024 (en prensa).
25. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. JAMA Health Forum. 2023;4:e231938.
26. Organización Mundial de la Salud (OMS). Informe sobre Inteligencia Artificial (IA) aplicada a la salud y seis principios rectores relativos a su concepción y utilización. (Consultado 18 Diciembre 2023). Disponible en: https://www.who.int/es/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use.
27. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with chatgpt and a call for papers. JMIR Med Educ. 2023;9:e46885.
28. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.
29. Chen A, Chen DO. Accuracy of chatbots in citing journal articles. JAMA Netw Open. 2023;6:e2327647.
30. Ryan DK, Maclean RH, Balston A, Scourfield A, Shah AD, Ross J. Artificial intelligence and machine learning for clinical pharmacology. Br J Clin Pharmacol. 2023 Oct 16. doi: 10.1111/bcp.15930.