Sociedad Neurológica Argentina
Filial de la Federación Mundial de Neurología

Neurología Argentina

www.elsevier.es/neurolarg

Artículo original

NeuroGPT, evaluando ChatGPT: Diagnóstico y tratamiento de 72 pacientes neurológicos

Alejandro Fernández Cabrera[a,*], *Jesús García de Soto*[b], *Paula Santamaría Montero*[a], *Héctor Chinea García*[c] y *Robustiano Pego Reigosa*[a]

[a] *Servicio de Neurología. Hospital Universitario Lucus Augusti, Lugo, España*
[b] *Unidad de doctorado en investigación médica, Universidad de Santiago de Compostela, Santiago de Compostela, España*
[c] *Facultad de Informática. Universidad de La Laguna, Santa Cruz de Tenerife, España*

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 10 de mayo de 2024
Aceptado el 9 de agosto de 2024
On-line el 7 de septiembre de 2024

Palabras clave:
Inteligencia artificial
ChatGPT
Neurología
Diagnóstico

RESUMEN

Introducción: Ha habido un auge significativo en el campo de la inteligencia artificial (IA) en los últimos años, especialmente en cuanto a accesibilidad y su uso en diferentes apartados. Este estudio intenta determinar si una IA puede diagnosticar pacientes de Neurología.
Objetivo: Evaluar la utilidad y precisión de *Chat Generative Pre-Trained Transformer* (ChatGPT) 3.5 como herramienta para realizar la historia clínica, diagnóstico y tratamiento en casos de patología neurológica.
Material y métodos: Se realizó un estudio observacional descriptivo cualitativo, sin intervención en pacientes, centrado en evaluar la utilidad y precisión de ChatGPT 3.5 para la toma de historia clínica, diagnóstico y tratamiento en pacientes con patología neurológica. La información proporcionada al neurólogo se introdujo en el modelo de lenguaje. Posteriormente, se realizó las preguntas que ChatGPT determinaba, y se suministró el examen neurológico completo. Se comprobó el diagnóstico de ChatGPT con el de dos neurólogos diferentes. El reclutamiento se realizó de mayo de 2022 a junio de 2023 en una consulta de Neurología de un hospital de tamaño mediano en España.
Resultados: Un total de 72 pacientes (edad mediana 58,71 años y 55,6% mujeres) fueron inscritos en este estudio. Las pruebas complementarias (PPCC) sugeridas por la IA se consideraron correctas en el 33,3% de los casos. La precisión en el diagnóstico de la IA fue del 44,4% y las recomendaciones de tratamiento fueron correctas en el 37,5%. El diagnóstico fue comprobado por dos neurólogos diferentes siguiendo las últimas guías nacionales e internacionales de Neurología. En la mayoría de los casos el diagnóstico entre ambos neurólogos coincidió con un coeficiente kappa de 0,94.

* *Autor para correspondencia.*
Correo electrónico: alejandrofcab@gmail.com (A. Fernández Cabrera).
https://doi.org/10.1016/j.neuarg.2024.08.002

Conclusiones: Aunque estamos en un momento de avance sin igual en el campo de la IA, no parece que en este momento ChatGPT pueda sustituir la valoración de un especialista en Neurología.

NeuroGPT, evaluating chatGPT: diagnosis and treatment of 72 patients

Keyword:
Artificial intelligence
ChatGPT
Neurology
Diagnosis

A B S T R A C T

Introduction: There has been a significant boom in the field of artificial intelligence in recent years, especially in terms of accessibility and its use in different areas. This study attempts to determine if an AI can diagnose neurology patients.

Objective: To evaluate the utility and accuracy of ChatGPT 3.5 as a tool for conducting patient history, diagnosis, and treatment in cases of neurological pathology.

Materials and methods: A descriptive qualitative observational study was conducted, without intervention in patients, focused on evaluating the utility and accuracy of ChatGPT 3.5 for taking patient history, diagnosis, and treatment in patients with neurological pathology. The information provided to the neurologist was entered into the language model. Subsequently, the questions determined by ChatGPT were asked, and the complete neurological examination was provided. ChatGPT's diagnosis was compared with that of two different neurologists. Recruitment took place from May 2022 to June 2023 in a neurology consultation at a medium-sized hospital in Spain.

Results: A total of 72 patients (median age 58.71 years and 55.6% female) were enrolled in this study. Complementary tests suggested by the AI were considered correct in 33.3% of cases. The accuracy of the AI's diagnosis was 44.4%, and treatment recommendations were correct in 37.5%. The diagnosis was checked by two different neurologists following the latest national and international Neurology guidelines. In most cases, the diagnosis between the two neurologists agreed, with a kappa coefficient of 0.94.

Conclusions: Although we are in an unprecedented era of advancement in the field of artificial intelligence, it does not seem that ChatGPT can currently replace the evaluation of a neurology specialist.

Introducción

Chat Generative Pre-Trained Transformer (ChatGPT) es un modelo de lenguaje conversacional basado en la inteligencia artificial (IA)[1]. Este modelo ha sido entrenado mediante el uso de bases de textos masivos en diferentes idiomas con el objetivo de generar respuestas a textos que introducimos[1]. Además, este modelo de lenguaje es ajustado de manera continua por técnicas de aprendizaje, tanto automatizadas como supervisadas y de refuerzo. Las instrucciones que introdujimos en la conversación se llaman *prompts*, y, para obtener una respuesta más precisa, debemos escribirlas de la forma menos genérica posible, añadiendo adjetivos o poniendo ejemplos del tipo de resultado que queremos obtener[2].

Desde la aparición de ChatGPT se han propuesto diferentes utilidades en múltiples campos, obteniendo resultados sorprendentes en diversos exámenes, tanto tipo test como escritos, incluyendo las pruebas de acceso a la licencia médica en EE. UU.[3]. Más controvertida resulta su utilidad en el ámbito médico, tanto clínico como investigador[4]. Se ha estudiado su aplicación, por ejemplo, a la hora de resolver dudas al médico o al paciente[5], realizar un diagnóstico radiológico[6] o escribir un artículo científico[7]. Recientemente incluso se ha realizado un estudio que comprobaba la precisión diagnóstica de ChatGPT para casos clínicos complejos con respuesta de opción múltiple, comprobado con respuestas de lectores de dicha revista[8].

A pesar de sus potenciales aplicaciones, esta IA también ha generado controversia, especialmente en lo relativo a la autoría de las publicaciones o ensayos, tanto científicos como académicos[9]. Además, en algunos casos, responde con información incorrecta o poco precisa, lo que podría limitar sus aplicaciones en la práctica clínica[4,5].

Objetivos

El objetivo de este estudio es valorar la precisión de ChatGPT 3.5 como herramienta para dirigir una anamnesis, realizar un diagnóstico y proponer el tratamiento de pacientes con patología neurológica.

Material y métodos

Se trata de un estudio observacional descriptivo cualitativo realizado en un hospital de segundo nivel.

Se reclutaron a los pacientes que acudían a consulta de Neurología de nuestro hospital desde noviembre de 2022 hasta mayo de 2023. Se consideraron e incluyeron algunos pacientes que se interconsultaban desde el servicio de urgencias del hospital a nuestro servicio. Se les ofreció participar en el estudio en el momento inicial de realizar la historia clínica. Todos los pacientes que aceptaron firmaron un consentimiento informado.

Se introducía la información dada por los pacientes en ChatGPT de la misma manera que la habían comunicado al médico. Realizábamos las preguntas a los pacientes, en tiempo real, que ChatGPT nos indicaba y se aportaba la respuesta del paciente a la IA. Posteriormente, indicábamos a ChatGPT la exploración neurológica realizada por el especialista. A continuación, se comprobaba qué pruebas complementarias (PPCC) solicitaría la IA y se ofrecían los resultados de estas a la IA. Finalmente, con toda esta información se le preguntaba sobre el diagnóstico más probable y su posible tratamiento.

Posteriormente se comparaba la información que proporcionaba el motor de lenguaje con la anamnesis, las PPCC, el diagnóstico y el tratamiento que un especialista en Neurología realizaba basándose en su propia experiencia y en las guías más actualizadas de las diversas patologías. El diagnóstico, las PPCC y el tratamiento se corroboraron con un segundo neurólogo ajeno, otorgándole la misma información que se había dado a ChatGPT. El diagnóstico realizado por los neurólogos se basó en las últimas guías clínicas de Neurología.

La introducción de información en ChatGPT se hizo con *prompts* adecuados y comprobados por un informático especialista en IA para evitar las respuestas vagas o imprecisas del motor de lenguaje. Debido al propio diseño de este estudio, donde dábamos la información al modelo de lenguaje tal cual nos la había aportado los pacientes, el diseño del *prompt* fue siempre el mismo, pero no el *prompt* en sí mismo.

Se consideraron las siguientes variables: edad, sexo, categoría de la patología neurológica (epilepsia, vascular, cefalea, demencias, trastornos del movimiento, neuromuscular, desmielinizante y otros), anamnesis, PPCC, diagnóstico y tratamiento.

En la categoría anamnesis, diagnóstico y tratamiento se consideraron los resultados «Correcto» – es decir, la IA realizó las mismas preguntas que hubiera realizado el neurólogo –, «Incorrecto» y «Aceptable» – si parte de la respuesta era correcta, pero faltaban preguntas que posteriormente sí que realizó el neurólogo y que resultaban cruciales para llegar al diagnóstico o elegir tratamiento o bien el diagnóstico o tratamiento era parcialmente correcto –.

Tabla 1 – Frecuencias y porcentaje de las subcategorías neurológicas

Categoría neurológica	Frecuencia	Porcentaje
Epilepsia	18	25%
Vascular	12	16,7%
Demencias	10	13,9%
Cefaleas	9	12,5%
Trastornos del movimiento	8	11,1%
Desmielinizante	7	9,7%
Neuromuscular	3	4,2%
Otros	5	6,9%

En la categoría PPCC se consideraron los siguientes resultados: «Correcto» e «Incorrecto». Dentro de la categoría de «Incorrecto» se diferenció entre «Exceso de pruebas» y «Defecto de pruebas». Se consideró que había un defecto de pruebas o un exceso de pruebas en base a las últimas guías de la sociedad española de Neurología o, en casos donde dichas guías no nombraran de manera específica cuándo solicitar una prueba, a consensos de expertos internacionales.

El análisis de datos se realizó tanto con dichos resultados como simplemente considerando «Correcto» e «Incorrecto», sumando los resultados de la categoría «Aceptable» a «Incorrecto».

Se realizó un análisis de todos los pacientes en total y posteriormente por subcategorías de los diferentes tipos de patología neurológica. El análisis estadístico se realizó con el programa SPSS® 25.0.

Resultados

Se le ofreció participar a un total de 121 pacientes y aceptaron participar 72 pacientes fueron incluidos en el estudio. El motivo de no aceptar fue siempre el mismo, desconfianza de dar «sus datos» a una IA o falta de tiempo. Cuatro pacientes no pudieron ser incluidos por presentar discordancia entre los neurólogos en el diagnóstico inicial. La edad media de los pacientes reclutados fue de 58,7 años con una desviación estándar de 21,6 (rango 19-91).

Hubo un cierto predominio de mujeres con un total de 55,6% (n = 40), el resto fueron hombres. La mayoría de los pacientes fueron vistos en consultas externas (69,4% n = 50) y el resto fueron interconsultas del servicio de urgencias a la guardia de Neurología.

En cuanto a la categoría de Neurología, la mayoría de los pacientes se encuadraban dentro de la categoría de epilepsia (25% n = 18), seguida por vascular (16,7% n = 12). El resto de los porcentajes puede verse en la tabla 1.

En un 40,3% (n = 23) la anamnesis realizada por la IA se consideró correcta, es decir, similar o igual a la que habría hecho un o una especialista en Neurología. Se consideró como aceptable en un 38,9% (n = 28). Falló únicamente en un 2,8% (n = 2), donde no realizó las preguntas adecuadas en ningún caso. En un 18,1% (n = 13) no se consideró la anamnesis puesto que no se permitió realizarla a la IA, por ejemplo, en un ictus con semiología de *total anterior circulation infarct* (TACI) donde la anamnesis es muy limitada por las propias características del paciente.

Tabla 2 – Anamnesis y pruebas complementarias en relación con la categoría neurológica

Categoría neurológica – anamnesis	Correcto	Aceptable	Fallo	No procede
Epilepsia	1 (5,5%)	10 (55,6%)	2 (11,1%)	5 (27,8%)
Vascular	3 (25%)	3 (25%)	0	6 (50%)
Cefalea	8 (88,9%)	1 (11,1%)	0	0
Demencias	5 (50%)	3 (30%)	0	2 (20%)
Trastornos del movimiento	5 (62,5%)	3 (37,5%)	0	0
Desmielinizante	4 (57,1%)	3 (42,9%)	0	0
Neuromuscular	2 (66,7%)	1 (33,3%)	0	0
Otros	1 (20%)	4 (80%)	0	0
Categoría neurológica – pruebas complementarias	**Correcto**	**Exceso**	**Defecto**	**Fallo**
Epilepsia	5 (27,8%)	9 (50%)	2 (11,1%)	2 (11,1%)
Vascular	8 (66,6%)	2 (16,7%)	1 (8,3%)	1 (8,3%)
Cefalea	4 (44,4%)	5 (55,6%)	0	0
Demencias	2 (20%)	5 (50%)	3 (30%)	0
Trastornos del movimiento	0	8 (100%)	0	0
Desmielinizante	3 (42,9%)	2 (28,6%)	2 (28,6%)	0
Neuromuscular	1 (33,3%)	2 (66,6%)	0	0
Otros	1 (20%)	3 (60%)	1 (20%)	0

Las PPCC que indicó la IA fueron las mismas que habría indicado un especialista en Neurología basándose en las guías y consensos de expertos en un 33,3% (n = 24) de los pacientes, mientras que falló en el resto. De los que se consideró que fallaba, en la gran mayoría (un 75% de estos) falló por «exceso» de pruebas, la mayoría de las veces siendo esta prueba una punción lumbar o una prueba de imagen como resonancia magnética cuando no estaban indicadas según las últimas guías en Neurología. En un 18,8% falló por defecto de pruebas, en el 6,3% restante se consideró que directamente fallaba porque indicaba pruebas que no estaban indicadas en la patología. Los resultados concretos por patologías pueden verse en la tabla 2.

En cuanto al diagnóstico la IA acertó el mismo en un 44,4% (n = 32) y falló en un 52,8% (n = 38). En los casos restantes no se le preguntó por diagnóstico porque se le dio directamente y se le preguntó por pruebas y tratamiento. El diagnóstico fue contrastado por el dado por los dos neurólogos, uno de ellos teniendo la misma información que se le había aportado a ChatGPT, el coeficiente Kappa entre el diagnóstico de ambos neurólogos fue de 0,94. En los cuatro casos donde el diagnóstico no coincidió entre neurólogos no se consideraron para el estudio.

En cuanto al tratamiento se consideró que el indicado por la IA era correcto en un 37,5% (n = 27) y que falló en un 55,6% (n = 40) de las veces. La mayoría de las veces que se consideró que la IA no elegía el tratamiento adecuado también había fallado en el diagnóstico (36 de las 40 veces), las cuatro veces restantes elegía un tratamiento que no venía indicado como primera opción o siquiera como opción en las últimas guías de Neurología. En el resto de los casos (n = 5, 6,9%) se consideró que el tratamiento que indicaba podía ser correcto, pero había alternativas mejores, por ejemplo, cuando recomendaba pautar un medicamento anticrisis, es decir, acertaba que el diagnóstico era de una crisis epiléptica y que debía ser tratada, pero no elegía el óptimo en base a las características del paciente, la mayoría de estos casos se consideró que no era óptimo por sugerir valproato en mujeres jóvenes en edad fértil como primera opción.

Tabla 3 – Diagnóstico en relación con la categoría neurológica

Categoría neurológica – diagnóstico	Correcto	Incorrecto	No procede
Epilepsia	2 (11.1%)	14 (77,8%)	2 (11,1%)
Vascular	8 (66,6%)	4 (33,3%)	0
Cefalea	8 (88,9%)	1 (11,1%)	0
Demencias	1 (10%)	9 (90%)	0
Trastornos del movimiento	7 (87,5%)	1 (12,5%)	0
Desmielinizante	3 (42,8%)	4 (57,2%)	0
Neuromuscular	3 (100%)	0	0
Otros	0	5 (100%)	0

Aunque el número de pacientes es pequeño también se intentó hacer un subanálisis por categorías neurológicas. Pueden verse los resultados en la tabla 2 y 3.

Llama la atención la aparente mayor capacidad para acertar en el diagnóstico en cefaleas y trastornos del movimiento, donde la IA acertó en un 88,9% (n = 8) y un 87,5% (n = 7) de las veces el diagnóstico respectivamente. Acierta un 100% (n = 3) de los pacientes en la categoría de neuromuscular. Falla en un 88,8% (n = 16) de los pacientes en la categoría de epilepsia y en un 90% (n = 9) de los pacientes en la categoría de demencia.

Relativo a la anamnesis por categorías apenas hay diferencias salvo en cefaleas donde se consideró correcta en un 88,9% (n = 8) de los pacientes en este subgrupo de patología y únicamente en un paciente se consideró «aceptable». El resto resulta menos destacable, siendo la categoría de «correcta» y «aceptable» similar en todas las subcategorías.

En cuanto a las PPCC por subgrupos llama la atención que siempre fallaba en la categoría trastornos del movimiento, lo cual contrasta con la alta fiabilidad en el diagnóstico que tenía. Prácticamente siempre fallaba en epilepsia, aunque en este caso es menos sorprendente puesto que la ratio diagnóstica que tenía en dicha categoría no era muy elevada.

No hubo prácticamente diferencias en cuanto al tratamiento por subgrupos, fallando y acertando de manera similar en todas las categorías salvo en cefaleas, donde acierta un 78% de las veces en el tratamiento. Una vez más,

en neuromuscular acierta el 100% de las veces y falla en un 83,3% de las veces en la categoría epilepsia.

Discusión

En cuanto a los resultados se puede comprobar como la IA tiende a solicitar más pruebas de las que solicitaríamos en vida real en un hospital. Hay que tener en cuenta que, si quizás no existieran las listas de espera o las PPCC no tuvieran coste, el planteamiento de «exceso de pruebas» de esta IA no sería considerado del todo incorrecto. Nos referimos con esto, lógicamente, al hecho de pruebas de imagen principalmente como RM y no a las veces que la IA recomendaba la realización de una punción lumbar cuando no estaba indicada.

Es bastante notorio el hecho de que el modelo de lenguaje sea capaz de "realizar" una anamnesis aceptable, haciendo las preguntas adecuadas al paciente, aunque pueda faltar algún matiz, prácticamente en un 80% de los casos. Este dato resulta sorprendente teniendo en cuenta que la anamnesis en Neurología es especialmente compleja.

En cuanto al diagnóstico por categorías tiene una gran capacidad de acertar en la categoría de vascular, aunque hay que tener en cuenta que todos los casos que acertó de esta subespecialidad fueron ictus de arteria cerebral media o lacunares donde, quizás, el diagnóstico es más directo y la clínica más clásica; los cuatro que falló eran ictus más complejos, uno de ellos talámico y tres con clínica menos común. Como curiosidad en este apartado, en uno de los casos en que falló la IA también erraron los dos neurólogos que hacían de control, llegando al diagnóstico mediante una resonancia magnética posterior. De similar manera acierta prácticamente en la totalidad de la categoría de trastornos de movimiento, aunque, como ya dijimos, solicita pruebas no necesarias, siendo la mayoría de los casos temblores esenciales, uno causado por fármacos y el resto por enfermedad de Parkinson. El único caso que falló el diagnóstico en esta categoría fue un temblor con características funcionales, lo cual si tenemos en cuenta las propias limitaciones del modelo de lenguaje resulta normal. Una vez más tenemos una limitación importante en el hecho de que tan pocos pacientes no son representativos de toda la patología que podemos encontrar en trastornos del movimiento.

Por último, cabe destacar la fiabilidad diagnóstica y de anamnesis en la categoría de cefaleas, donde ChatGPT acertaba casi siempre, tanto en anamnesis como en diagnóstico y tratamiento. Una vez más, todos los pacientes tenían migraña o cefalea tensional, siendo el único error una cefalea menos común, una neuralgia supratroclear. En este caso, ChatGPT sugería que podía tratarse de una cefalea «atípica» o «daño de un nervio» pero no encontraba el diagnóstico concreto ni su tratamiento.

En la categoría de demencias falla prácticamente siempre, y en todos los casos por «exceso», diagnosticando de probable demencia o deterioro cognitivo a pacientes que realmente presentaban pérdidas de memoria relacionadas con la edad sin datos de demencia.

En la categoría de epilepsia es donde encontramos más datos. Esto se debe a un sesgo de selección puesto que el investigador principal que reclutaba se dedica principalmente a epilepsia. Asimismo, en esta categoría encontramos una gran proporción de anamnesis únicamente «aceptables». Esto podría estar motivado una vez más por el hecho de que el neurólogo podía ser más estricto a la hora de valorar las respuestas de la IA.

Hay que matizar que en la categoría de neuromuscular el número de pacientes es excesivamente pequeño y en los tres casos el diagnóstico fue el mismo: una polineuropatía tóxico-metabólica (diabética), por lo que el hecho de que acertase el 100% probablemente no sea muy significativo. La complejidad que tienen los pacientes con patología neuromuscular haría que, en caso de tener trastornos más variados, la IA posiblemente fallara muchísimo más. Hacen falta más pacientes y estudios en ese aspecto para comprobar esta fiabilidad.

Un factor importante es la dificultad que existe para que ChatGPT suministre la información deseada, es decir, es difícil redactar los *prompts* necesarios, de los que ya hablamos en la introducción, para obtener los resultados esperados. La mayoría de las veces que se pregunta a ChatGPT sobre medicina, intenta, posiblemente con buen criterio, evadir la respuesta y recomendarte acudir a un especialista en Neurología. Se necesita cierta destreza a la hora de «*preguntarle*» a esta IA para que responda lo que, aparentemente, sabe. En ese aspecto contar con ayuda de un informático especialista en IA fue vital para este estudio.

En el momento de escribir este artículo no hemos encontrado ninguno similar a nuestro estudio, por lo que resulta difícil establecer comparaciones. Hay un estudio realizado por especialistas en gastroenterología que analizaron la capacidad de ChatGPT de responder preguntas sobre hepatología. En ese estudio[5] ChatGPT respondió correctamente en aproximadamente un 80% de las preguntas sobre cirrosis y en un 75% aproximadamente sobre hepatocarcinoma, aunque consideraron que sólo un 47.3% y un 41.1% respectivamente fueron respuestas «completas».

Como limitaciones a este estudio podemos nombrar primero el número de pacientes, que es limitado dada la gran variabilidad de patología neurológica. Asimismo, aunque es lógico que el estudio se ha realizado únicamente en pacientes con patología neurológica, existe un sesgo de selección, además de la alta proporción de pacientes con epilepsia dado que el investigador principal tiene una mayoría de pacientes con epilepsia. Por último, por la propia naturaleza del estudio, donde administramos la propia información de los pacientes a la IA, esta no siempre contó con la misma información en todos los pacientes.

Son necesarios nuevos estudios donde se pueda evaluar de manera objetiva el acierto de la IA valorando la patología neurológica, quizás centrados únicamente en un tipo de patología, por otro lado, sería recomendable realizar nuevamente los estudios con la nueva versión de ChatGPT.

Conclusiones

Aunque estamos en un momento de avance sin igual en el campo de la IA, no parece que en este momento ChatGPT pueda sustituir la valoración de un especialista en Neurología. Prácticamente en tres de cada cuatro casos consideró pruebas que no estaban indicadas.

Hacen falta más estudios para comprobar si se produce una mejoría con versiones futuras, patologías diferentes o para valorar si pudiera usarse como apoyo al neurólogo o médicos de otras especialidades.

Financiación

Este artículo no ha recibido financiación pública ni privada de ningún tipo.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

BIBLIOGRAFÍA

1. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.
2. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 2023;25:e50638.
3. 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 (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.
4. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare. 2023;11:887.
5. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023;29:721–32.
6. Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. J Am Coll Radiol. 2023;20:990–7.
7. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health. 2023;13:01003.
8. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to Diagnose Complex Clinical Cases. NEJM AI. 2023;1:1–3.
9. Thorp HH. ChatGPT is fun, but not an author. Science (1979). 2023;379:313.