revistadelta

https://dx.doi.org/10.1590/1678-460X202541368342

Artículos

Inteligencia artificial para mejorar la comprensión de informes médicos: un análisis comparativo de legibilidad

Using artificial intelligence to enhance comprehension of medical reports: a comparative readability analysis

Inteligência artificial para melhorar a compreensão de relatórios médicos: uma análise comparativa de legibilidade

Adéla Koťátková*
Irene Peinado Zanón**
María Álamo Rodríguez***
Alejandra López Hernández****
Emma Plana*****
Manuel Miralles Hernández****** [1]

RESUMEN

La comunicación sigue siendo uno de los principales retos en la relación médico-paciente. La dificultad de los pacientes para comprender los informes médicos afecta directamente su capacidad de tomar decisiones informadas. El objetivo de este estudio es analizar la capacidad de ChatGPT para generar

1. Facultat de Medicina, Universitat de València (España).
* Departament de Filologia i Cultures Europees / Institut Interuniversitari López Piñero, Universitat Jaume I, Castelló de la Plana (España). https://orcid.org/0000-0003-2395-7473. E-mail: adela.kotatkova@uji.es.
** Servicio de Angiología y Cirugía Vascular, Hospital Universitari i Politècnic La Fe, Valencia (España). https://orcid.org/0009-0001-5925-8618. E-mail: peinado_irezan@gva.es.
*** Servicio de Angiología y Cirugía Vascular, Hospital Universitari i Politècnic La Fe, Valencia (España). https://orcid.org/0009-0006-2637-810X. E-mail: alamo_mar@gva.es.
**** Facultat de Medicina, Universitat de València (España). https://orcid.org/0009-0004-2931-529X. E-mail: lopez_aleher@gva.es.
***** Grupo acreditado de Hemostasia, Trombosis, Arteriosclerosis y Biología Vascular, Instituto de Investigación Sanitaria (IIS) La Fe, Valencia (España). https://orcid.org/0000-0002-8457-6984. E-mail: emma_plana@iislafe.es.
****** Servicio de Angiología y Cirugía Vascular, Hospital Universitari i Politècnic La Fe, Valencia (España). https://orcid.org/0000-0001-8340-8172. E-mail: miralles_manher@gva.es.

versiones simplificadas de informes de alta hospitalaria y evaluar tanto su comprensibilidad como su precisión técnica. A través del uso de cuatro fórmulas de legibilidad (índice de lecturabilidad de Fernández-Huerta (IL), perspicuidad de Szigriszt-Pazos (IP), escala Inflesz (EI) e índice de legibilidad μ (Iμ)) y un análisis estadístico comparativo no paramétrico, se evidenció que los informes adaptados por ChatGPT mejoraron su comprensibilidad en un 27% en comparación con los originales, reduciendo significativamente el nivel de dificultad de los textos. La precisión del lenguaje médico (errores/1000 palabras) fue evaluada por un panel de expertos médicos y filólogos. Las principales fuentes de imprecisión identificadas fueron: terminología (2,1‰), medicación (1,6‰), abreviaciones (0,92‰) y las denominadas «alucinaciones» del chatbot (0,18‰). A pesar de que la tasa de errores es baja, sigue siendo una limitación importante.

Palabras clave: *inteligencia artificial; ChatGPT; informe de alta; lenguaje médico; legibilidad*

ABSTRACT

Communication remains one of the key challenges in the doctor-patient relationship, as patients' difficulty in understanding medical reports directly impacts their ability to make informed decisions. The aim of this study is to analyze ChatGPT's ability to generate simplified versions of hospital discharge reports and evaluate both their comprehensibility and technical accuracy. Using four readability formulas (Fernández-Huerta readability index (IL), Szigriszt-Pazos perspicuity index (IP), Inflesz scale (EI), and μ readability index (Iμ)) and a non-parametric comparative statistical analysis, the findings showed that ChatGPT-adapted reports improved comprehensibility by 27% compared to the original texts, significantly reducing their level of difficulty. Medical language accuracy (errors per 1,000 words) was evaluated by a panel of medical experts and linguists. The main sources of inaccuracy were identified as: terminology (2.1‰), medication (1.6‰), abbreviations (0.92‰), and chatbot hallucinations (0.18‰). Despite the low error rate, it remains an important limitation.

Keywords: *artificial intelligence; ChatGPT; discharge report; medical language; readability*

RESUMO

A comunicação continua sendo um dos principais desafios na relação médico-paciente. A dificuldade dos pacientes em compreender os relatórios médicos afeta diretamente sua capacidade de tomar decisões informadas. O objetivo

revistadelta
41.3
2025

deste estudo é analisar a capacidade do ChatGPT de gerar versões simplificadas de relatórios de alta hospitalar e avaliar tanto sua compreensibilidade quanto sua precisão técnica. Utilizando quatro fórmulas de legibilidade — índice de Fernández-Huerta (IL), perspicuidade de Szigriszt-Pazos (IP), escala Inflesz (EI) e índice de legibilidade μ (Iμ) — e uma análise estatística comparativa não paramétrica, observou-se que os relatórios adaptados pelo ChatGPT melhoraram sua compreensibilidade em 27% em comparação com os textos originais, reduzindo significativamente o nível de dificuldade. A precisão da linguagem médica (erros/1000 palavras) foi avaliada por um painel de especialistas médicos e filólogos. As principais fontes de imprecisão identificadas foram: terminologia (2,1‰), medicação (1,6‰), abreviações (0,92‰) e "alucinações" do chatbot (0,18‰). Apesar da baixa taxa de erros, esta permanece uma limitação importante da ferramenta.

Palavras-chave: *inteligência artificial; ChatGPT; relatório de alta; linguagem médica; legibilidade*

1. Introducción

En el enfoque actual de la medicina, el paciente no solo debería estar en el centro de las decisiones tomadas, sino que también debería poder participar activamente en ellas. Sin embargo, esto solo es factible si se le proporciona información que sea comprensible para él. De este modo, el paciente debería poder interpretar, entre otros aspectos, qué le está sucediendo, qué pruebas se le han realizado, cuáles son sus resultados y cuáles son las opciones de tratamiento o seguimiento. Por eso, una adecuada comunicación siempre debe ser uno de los pilares fundamentales de la relación entre el médico y el paciente. En el ámbito de la comunicación sanitaria, generalmente se suele poner más énfasis en la modalidad oral, que convencionalmente establece la relación entre el profesional y el paciente, mientras que la comunicación escrita a menudo se descuida o se relega a un segundo plano.

Este estudio se centra en la comunicación escrita, ya que gran parte de la información relevante para los pacientes se plasma en los diferentes textos médicos. Si bien estos documentos incluyen todos los datos pertinentes sobre el paciente, su contenido y su forma no suelen estar diseñados para su comprensión por un lector no especializado.

Investigaciones recientes destacan que los pacientes bien informados tienden a adherirse más diligentemente a las recomendaciones de los

profesionales sanitarios y muestran una mayor disposición a seguir protocolos y a participar activamente en su tratamiento (Gómez-González et al., 2013). Por el contrario, la falta de comprensión de los documentos proporcionados por los profesionales de la salud puede generar desconfianza en el sistema sanitario y fomentar la búsqueda de informaciones en línea, a menudo con datos falsos y potencialmente peligrosos.

Así pues, se hace evidente la necesidad de disponer de documentos escritos específicamente diseñados para los pacientes. Sin embargo, la tarea de reescribir informes requiere un tiempo considerable y, probablemente, sería insostenible en el sistema sanitario actual. En este contexto, recientemente han surgido herramientas basadas en la inteligencia artificial (IA), como los chatbots, que ofrecen nuevas oportunidades para trabajar con textos médicos.

Los chatbots, incluido ChatGPT, el más conocido en Europa, no solo tienen la capacidad de generar textos, aspecto que ha recibido la mayor atención, sino también la de procesar documentos complejos como los informes médicos, en tiempos inimaginables para los seres humanos. Entre sus capacidades se encuentran las de resumir, traducir o, por ejemplo, adaptar contenidos a públicos menos especializados. Precisamente, esta última abre una nueva vía en relación con los informes médicos. Con indicaciones adecuadas (*prompts*) y un proceso previo de pruebas, un chatbot podría reescribir un informe para adaptarlo al nivel de comprensión del paciente, por ejemplo, explicando los términos médicos o desplegando las tan frecuentes siglas.

Por otra parte, para evaluar en qué medida una herramienta mejora la comprensibilidad de los informes, se pueden aplicar diversos índices de legibilidad. Este concepto se define como la facilidad con la que un texto puede ser leído y comprendido, teniendo en cuenta elementos tipográficos, presentación en la página, estilo, claridad en la exposición, la forma de escribir y el lenguaje en sí. En este trabajo, nos centramos específicamente en el aspecto de la legibilidad que analiza la estructura del texto, incluyendo la longitud de las palabras, las frases, y la construcción gramatical; en resumen, lo que se conoce como legibilidad lingüística formal (Barrio Cantalejo & Simón Lorda, 2003).

La legibilidad se puede evaluar con herramientas automáticas que aplican fórmulas matemáticas para medir la dificultad de lectura y comprensión de los textos. Trabajos previos centrados en la legibilidad de

revistadelta
41.3
2025

los textos médicos como los consentimientos informados (Rubiera et al., 2004; Ramírez-Puerta et al., 2013), las páginas web sobre cuestiones de salud (Blanco Pérez & Gutiérrez Couto, 2002) y los materiales educativos en salud para usuarios migrantes (Pelicano-Piris, 2022) han demostrado la aplicabilidad de estas herramientas. También han confirmado objetivamente que los documentos médicos tienden a ser difíciles de leer y entender para el ciudadano medio y, por ello, no siempre cumplen la finalidad básica para la que fueron redactados.

En este contexto, el objetivo de este estudio es doble: por un lado, detectar qué elementos hay que adaptar al lenguaje del paciente, desarrollando un *prompt* preciso para reescribir los informes de alta hospitalaria usando ChatGPT; por otro, analizar la legibilidad de los informes originales del Servicio de Angiología y Cirugía Vascular del Hospital Universitari i Politècnic La Fe de Valencia (España), para luego compararla con la de los informes adaptados al paciente a través de ChatGPT.

2. Metodología

Para realizar este estudio se escogió una muestra representativa de informes de alta hospitalaria de los principales procedimientos realizados en el Servicio de Angiología y Cirugía Vascular del Hospital Universitari i Politècnic La Fe.

Tomando como referencia un nivel de legibilidad y comprensibilidad de los informes originales inferior al 50% (43,9-54,7%) clasificados como *difícil* o *bastante difícil* según los parámetros utilizados, y el incremento necesario para alcanzar el estrato de *normal* o *bastante fácil*, estimamos que sería necesario aumentarlo en al menos un 15-20% sobre dicho umbral para que su eficacia resultara clínicamente relevante.

Por tanto, se estimó que se requerirían al menos 10 informes de cada procedimiento para detectar una diferencia igual o superior al 18%, con un riesgo alfa de 0,05 y beta de 0,2 en un contraste bilateral. Se asume una desviación estándar (DE) del 20% y la ausencia de pérdidas entre ambas mediciones. Para el cálculo del tamaño muestral, se utilizó el software Granmo (versión 7.11; Institut Municipal d'Investigació Mèdica, Barcelona).

Bajo esta premisa, se seleccionaron 10 informes de alta de forma aleatorizada sobre los últimos 100 pacientes ingresados en proporción 1:10

por cada uno de los 5 procedimientos principales que se llevan a cabo en este servicio: endarterectomía carotídea, cirugía de varices, acceso vascular para hemodiálisis, reparación abierta de aneurisma de aorta y reparación endovascular de aneurisma de aorta.

Estos 50 informes originales fueron anonimizados y procesados por ChatGPT (modelo GPT-3.5[2], versión de marzo de 2024), utilizando un *prompt* en desarrollo, como parte de una colaboración con el Departament de Filologia i Cultures Europees de la Universitat Jaume I. Como resultado de este proceso, se generaron 50 informes adaptados para los pacientes.

Se realizó una evaluación lingüística cualitativa detallada de los informes médicos originales, teniendo en cuenta estudios referentes como el de Estopà & Domènech-Bagaria (2019), para identificar los elementos que presentan mayores dificultades de comprensión para los pacientes. El objetivo del análisis fue diseñar un *prompt* para ChatGPT que enfocara la atención en los elementos que había que adaptar, simplificar o explicar para que fueran accesibles a un público lo más amplio posible.

Para evaluar y comparar el grado de comprensión de los informes originales y los adaptados por ChatGPT se utilizaron índices de legibilidad basados en algoritmos específicos para el español. Estos índices se obtuvieron de forma automática con una herramienta sencilla para su cálculo, disponible en la página web www.legible.es (Muñoz Fernández, 2018). En concreto, para este estudio se usaron 4 índices: 1) escala de lecturabilidad de Fernández-Huerta (1959); 2) nivel de perspicuidad de Szigriszt-Pazos (1993); 3) escala Inflesz, reinterpretación del anterior (Barrio et al., 2008); 4) índice de legibilidad µ (Muñoz & Muñoz, 2006).

La escala de Fernández Huerta (1959) sirve para medir la lecturabilidad, es decir, la dificultad para comprender un texto, y se basa en el promedio de sílabas por palabra y la media de palabras por frase. Su fórmula se expresa como: L= 206.84 – 0.60P – 1,02F (P: promedio de sílabas por palabra; F: media de palabras por frase). Asigna puntuaciones a los textos en una escala del 0 al 100, categorizándolos en 7 niveles educativos.

El índice de Szigriszt-Pazos mide la perspicuidad del texto, es decir, el nivel de compresión del estilo con el que se escribe. Su fórmula es:

2. Dado el rápido desarrollo de los modelos de lenguaje, es posible que versiones más recientes de GPT (como GPT-4 o posteriores) y otros LLMs, como Gemini, Claude o Llama, ofrezcan mejoras adicionales en legibilidad o presenten diferentes patrones de error.

revistadelta
41.3
2025

P=206.835 – 62.3S/P – P/F (S: total de sílabas; P: cantidad de palabras; F: número de frases). También tiene un rango del 0 al 100, donde cada intervalo equivale a un tipo de publicación y un nivel de estudios.

La escala INFLESZ se calcula como el índice Szigriszt-Pazos, pero introduce una reinterpretación de los resultados para el lector español medio. Esta escala está validada para evaluar la legibilidad de textos dirigidos a pacientes y, por lo tanto, se ha usado en numerosos estudios para valorar el índice de comprensión de diferentes textos médicos dirigidos a la población sin formación específica.

Finalmente, el índice de legibilidad μ calcula la facilidad de lectura de un texto a partir del número de palabras, la media y la varianza del número de letras de las palabras. Su fórmula es: $\mu = (n/n-1)(\bar{x}/\sigma^2) \times 100$ (n: número de palabras; $\bar{x}$: media del número de letras por palabra; σ^2: su varianza).

Para evaluar la precisión del lenguaje médico introducido por el chatbot, se examinó una muestra aleatoria de 10 informes (2 por cada procedimiento) por un equipo mixto de expertos médicos (2 especialistas en Angiología y Cirugía Vascular) y filólogos (2 doctores en Lingüística). Todos los textos fueron examinados de forma ciega e independiente en la búsqueda de los errores tipificados en las distintas categorías. Se escogieron 4 parámetros representativos para valorar la adecuada transmisión de la información relevante para el paciente y objetivar los fallos en ésta: terminología, abreviaciones, alucinaciones (información falsa generada por el chatbot) y medicación. Tras revisar los errores identificados por los distintos miembros del equipo, estos fueron confirmados por consenso, prevaleciendo, en caso de duda, la opinión del cirujano senior en caso de terminología médica y del filólogo/a más experto/a en caso de errores gramaticales o semánticos. A cada error de cada categoría se le asignó un valor, para poder contabilizarlo, matizando los errores completos (1) y los errores parciales o imprecisiones (0,5) en relación al número total de palabras de los informes originales. Después, ese valor se convirtió a su equivalencia como tasa por cada 1000 palabras. en función de los parámetros expuestos anteriormente.

Se realizó un análisis descriptivo de la población (informes de alta): media +/- desviación estándar de las variables cuantitativas (parámetros lingüísticos analizados) y proporciones de las cualitativas. Ante el reducido tamaño de la muestra analizada, y comprobar la falta de homogeneidad de la varianza en la prueba de Levene, se optó por el uso de pruebas no paramétricas para el análisis estadístico de los resultados. Se llevó a cabo un

estudio de comparación de medias entre los resultados del informe inicial y tras su reedición por el chatbot (utilizando la prueba de Wilcoxon para medidas repetidas) y entre los distintos tipos de informe (mediante la prueba de Kruskal-Wallis para grupos independientes). Para el análisis se utilizó el software SPSS (IBM Statistical Package for the Social Sciences, versión 27.0), considerando significativo un valor de $p<0,05$.

El estudio se llevó a cabo según la declaración de Helsinki con aprobación de sendos comités (ético y científico) del Hospital Universitario y Politécnico La Fe (referencia: 2024-229-1).

3. Resultados

Principales problemas de comprensión para los pacientes

En el análisis lingüístico se han identificado seis áreas principales que presentan mayores dificultades de comprensión para los pacientes (Tabla 1): la terminología médica, las siglas y otras abreviaciones, los nombres de medicamentos, las expresiones típicas del personal sanitario (pueden herir a los pacientes o dañar la relación médico-paciente), la sintaxis confusa (en forma de notas) y el exceso de datos.

Tabla 1 – Comparativa de la complejidad de las informaciones médicas entre los informes médicos originales y los reescritos mediante ChatGPT.

	Informe original	**Informe adaptado con ChatGPT**
Terminología médica	- Aneurisma de aorta abdominal infrarrenal - Vértigo e hipoacusia derecha. - Hernia inguinal bilateral - Herpes zoster dorsal izquierdo 2016	Aneurisma de aorta abdominal infrarrenal (dilatación en la parte de la aorta que está debajo de los riñones). Vértigo e hipoacusia derecha (mareos y pérdida de audición en el lado derecho) Hernia inguinal bilateral (salida del intestino a través de un punto débil en la pared del abdomen en ambos lados). Herpes zoster dorsal izquierdo en 2016 (una infección viral que causa dolor y ampollas en la piel).
Abreviaciones	Antecedentes No RAM FRCV: -HTA -No DM. -DL.	Antecedentes No tiene reacciones adversas a medicamentos FRCV: Factores de Riesgo Cardiovascular (condiciones o hábitos que pueden aumentar el riesgo de desarrollar enfermedades del corazón): Tiene presión arterial alta (HTA). No tiene diabetes (No DM). Padece de niveles altos de colesterol o grasas en la sangre (DL).

revistadelta
41.3
2025

Nombres de medicamentos	Tratamiento habitual: -CANDESARTAN 16 MG -ROSUVASTATINA VIATRIS 10MG -ACETILSALICILICO ACIDO 100 MG -BILASTINA 20 MG -DEPRAX 100MG -SYMBICORT FORTE TURBUHALER 320/9MCG -TAMSULOSINA 0,4 MG -PANTOPRAZOL 20 MG.	Tratamiento habitual: Candesartán 16 mg (baja la presión arterial). Rosuvastatina Viatris 10 mg (controla el colesterol). Ácido acetilsalicílico 100 mg (aspirina, se usa para prevenir coágulos sanguíneos). Bilastina 20 mg (antihistamínico, para alergias). Deprax 100 mg (antidepresivo). Symbicort Forte Turbuhaler 320/9 mcg (para los pulmones). Tamsulosina 0.4 mg (para la próstata). Pantoprazol 20 mg (para el estómago: reduce ácido).
Expresiones típicas del personal sanitario	No RAM conocidas (en orion Pregabalina, pero el paciente niega). Niega tabaquismo y consumo enólico.	No ha tenido reacciones adversas a medicamentos (significa que no ha tenido problemas graves debido a los medicamentos). No bebe alcohol.
Sintaxis confusa	Evolución clínica satisfactoria tras stent carotídeo derecho, sin complicaciones inmediatas. Subjetivamente se encuentra bien, sin focalidad neurológica. Sitio de punción inguinal derecha sin complicaciones, no hematoma ni masas pulsátiles. Analítica de control postoperatorio sin alteraciones relevantes.	Le colocaron un tubo metálico llamado “stent” en la arteria carótida derecha para mantenerla abierta. Se siente bien y no tiene problemas neurológicos. La zona de punción en la ingle derecha está bien, sin complicaciones. Sus análisis de sangre están normales.
Exceso de datos	BIOQUIMICA: GLUCOSA: 103 mg/dL; UREA: 57 mg/dL; CREATININA: 1,01 mg/dL; FILTRADO GLOMERULAR (CKD-EPI): 65 mL/min/1,73 m2; PROTEINA C REACTIVA: < 1,0 mg/L; CLORO: 106 mEq/L; SODIO: 141 mEq/L; POTASIO: 4,3 mEq/L; HEMOGRAMA: HEMATIES: 4,39 x 10^6/µL; HEMOGLOBINA: 13,5 g/dL; HEMATOCRITO: 41,6 %; VOLUMEN CORPUSCULAR MEDIO: 95 fL; HBG CORPUSCULAR MEDIA: 30,8 pg; CONC. HBG CORPUSCULAR MEDIA: 32,5 g/dL; RDW-SD: 42,7 fL; RDW-CV: 12,4 %; ERITROBLASTOS %: 0,0 %; LEUCOCITOS: 7,67 x 10^3/µL; NEUTROFILOS: 5,47 x 10^3/µL; LINFOCITOS: 1,34 x 10^3/µL; MONOCITOS: 0,67 x 10^3/µL; EOSINOFILOS : 0,16 x 10^3/µL; BASOFILOS: 0,03 x 10^3/µL; NEUTROFILOS %: 71,3 %; LINFOCITOS %: 17,5 %; MONOCITOS %: 8,7 %; EOSINOFILOS %: 2,1 %; BASOFILOS %: 0,4 %; PLAQUETAS: 154 x 10^3/µL; MPV: 9,5 fL; PDW: 10,6 fL; P-LCR: 21,4 %;	Análisis de sangre muestra valores dentro de rangos normales.

RAM: reacciones adversas a medicamentos; FRCV: factores de riesgo cardiovascular; HTA: hipertensión arterial; DM: diabetes mellitus; DL: dislipemia

Comparación de la legibilidad de los informes originales y los adaptados con ChatGPT

Los resultados obtenidos tras la realización del análisis lingüístico y estadístico se recogieron de manera independiente, tal y como se muestra en la Tabla 2. Cada una de las columnas representa la comparación entre informes realizados por ChatGPT y los informes originales (IO) para cada uno de los cinco procedimientos quirúrgicos. Se evidencia una mejora estadísticamente significativa en la puntuación de todas las escalas de los informes adaptados por ChatGPT respecto a los informes de alta originales.

Tabla 2 – Resultados de legibilidad por procedimiento. Variación entre el valor promedio del informe original (IO) y el adaptado (ChatGPT)

	Endarterectomía carotídea			Fleboextracción			Reparación abierta de aneurisma aórtico			Acceso de hemodiálisis			Reparación endovascular de aneurisma aórtico		
	IO	Chat GPT	P-valor	IO	Chat GPT	P-valor	IO	Chat GPT	P-valor	IO	Chat GPT	P-valor	IO	Chat GPT	P-valor
Lecturabilidad (Escala Fernández Huerta)	49,549	63,879	0,007	54,696	71,412	0,005	49,784	62,016	0,009	53,688	64,774	0,007	48,913	63,074	0,005
DE	5,18	5,80		8,37	3,49		3,94	5,52		4,86	6,45		2,42		
Nivel de perspicuidad (I. Szigriszt-Pazos)	44,066	58,978	0,005	49,283	66,565	0,005	44,384	56,820	0,005	48,252	61,452	0,005	43,895	58,019	0,005
DE	5,33	6,03		8,57	3,60		4,00	5,72		5,1	5,32		3,08		
Escala de legibilidad INFLESZ (Barrio)	44,066	58,978	0,005	49,283	66,565	0,005	44,384	56,820	0,005	48,252	61,452	0,005	43,895	58,019	0,005
DE	5,33	6,03		8,57	3,60		4,00	5,72		5,1	5,32		3,08		
Facilidad lectora (legibilidad µ)	45,989	54,673	0,005	46,322	54,943	0,005	45,533	51,296	0,005	46,137	60,937	0,005	45,752	51,824	0,005
DE	2,79	5,06		3,58	4,49		2,12	3,65		3,61	8,16		1,53		

En la Tabla 3 se presentan las diferencias encontradas en los 50 informes analizados, sin distinguir entre los diferentes procedimientos quirúrgicos. Se evidenció una mejora estadísticamente significativa en todos los parámetros evaluados, con un aumento global en sus valores del 27 +/- 5,7%, oscilando entre el 19% (para la legibilidad media) y el 31% (para Szigriszt-Pazos e Inflesz). La estratificación de las puntuaciones obtenidas en los diversos

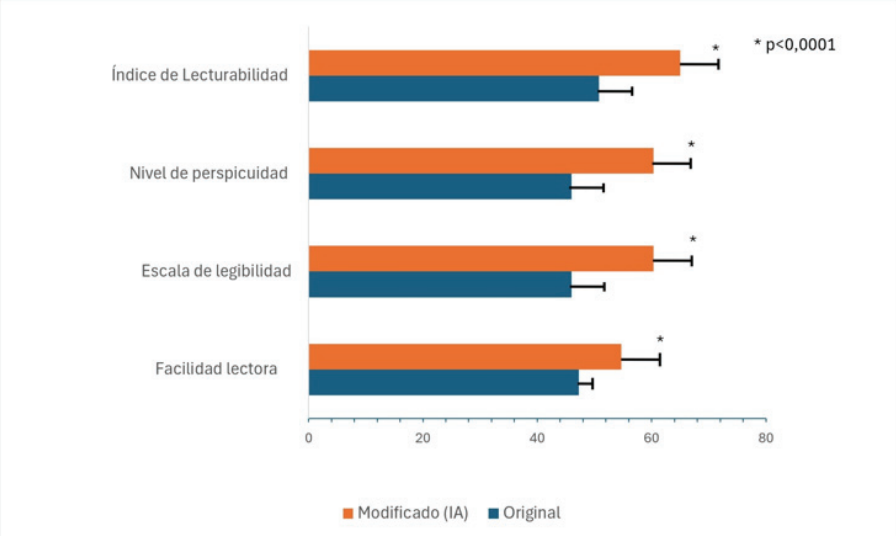
revistadelta
41.3
2025

parámetros revela un nivel de legibilidad clasificado como *difícil/bastante difícil* para los informes originales, mientras que para los informes adaptados se clasificó como *normal/un poco difícil* (Figura 1).

Tabla 3 – Resultados globales de comparación entre informes originales y adaptados

	Lecturabilidad (Escala Fernández Huerta)	Nivel de perspicuidad (I. Szigriszt-Pazos)	Escala de legibilidad INFLESZ (Barrio)	Facilidad lectora (legibilidad μ)
Original	50,80 ± 5,64	45,97 ± 5,79	45,97 ± 5,79	47,28 ± 2,74
Valoración según escala	difícil/algo difícil	bastante difícil	algo difícil	difícil
ChatGPT	65,031 ± 6,53	60,36 ± 6,57	60,36 ± 6,57	54,73 ± 6,15
Valoración según escala	normal	normal	normal	un poco difícil
Diferencia	27%	31%	31%	19%
P valor	<0,0001	<0,0001	<0,0001	<0,0001

Figura 1 – Mejora de los diferentes índices de legibilidad de los informes de alta (n=50) tras ser reeditados con inteligencia artificial. Con pequeñas variaciones, dependiendo del parámetro, un valor superior a 50 se considera como una disminución en el nivel de dificultad, pasando de *difícil/muy difícil* a *normal/no muy difícil*.



Desde el punto de vista médico, el análisis de precisión clínica ofrecida por el chatbot demostró que las principales fuentes de inexactitud fueron: terminología, medicación (incluyendo dosis/unidades), abreviaciones y alucinaciones, con una tasa de 2,1; 1,6; 0,92 y 0,18 por cada 1000 palabras, respectivamente. Los datos de origen de este análisis aparecen reflejados en la Tabla 4.

Tabla 4 – Principales fuentes de inexactitud de los informes adaptados por IA

	HD1	HD2	V1	V2	C1	C2	AB1	AB2	EV1	EV2	TOTAL
N.º palabras	108	411	357	321	509	699	911	688	698	730	5432
Terminología	1,5	1,5	1	1,5	1	1	0,5	2	1,5	0	11,5
Tratamiento	0	0,5	1	1	1	2	1	0	1	1	8,5
Alucinaciones	0	0	0	0	0	0	0	0	1	0	1
Abreviaciones	0	0	0	0	1	0	0	1	1,5	1,5	5

N: número; HD: acceso para hemodiálisis; V: cirugía de varices; C: endarterectomía carotídea; AB: reparación abierta de aneurisma de aorta; EV: reparación endovascular de aneurisma de aorta. Se analizaron dos informes por cada patología.

4. Discusión

La necesidad de asegurar la accesibilidad de los documentos médicos para la ciudadanía ha llevado al uso de índices que permitan evaluar de manera objetiva su legibilidad. Simón et al. (1996) fueron pioneros en este campo. Analizaron la legibilidad de 16 formularios de consentimiento informado, encontrando que solo 5 eran accesibles para el ciudadano medio. Resulta notable que los autores no se sorprendieran por los bajos resultados, considerándolos un reflejo de la perspectiva paternalista del rol médico de aquel tiempo. Por consiguiente, que nuestro estudio presente mejores resultados podría interpretarse como una tendencia positiva en la comunicación médica.

Esta línea de investigación continuó con Barrio et al. (2008) , creando el índice Inflesz, a partir de la revisión y el análisis de 210 publicaciones, incluyendo prensa popular y revistas científicas. En dicho estudio, se determinó que los textos con una puntuación superior a 55 puntos en el índice Inflesz se consideran accesibles, lo cual estableció un estándar de comprensibilidad. Este índice ha ganado reconocimiento por su eficacia, evaluando la legibilidad de textos médicos especializados para el lector

revistadelta
41.3
2025

español medio, y se ha consolidado como referente en este campo. En nuestro estudio, la aplicación de los índices Inflesz y Szigriszt-Pazos fue particularmente reveladora, mostrando una diferencia del 31% entre los informes originales y los adaptados. Aunque el incremento de la comprensibilidad de los informes reescritos por ChatGPT resultó ser menor de lo esperado, dado que solo los informes de cirugía de varices llegaron a considerarse *algo fáciles*, la mejora sigue siendo significativa para todos los informes adaptados. De hecho, la mayoría de estos obtuvieron puntuaciones que los sitúan en la categoría de *normal*, según todas las escalas de legibilidad empleadas en el estudio.

La legibilidad de los informes de alta hospitalaria en español, objeto de nuestra investigación, ha sido evaluado previamente por Estopà et al. (2020) y Porras-Garzón & Estopà (2020). Este trabajo es de particular interés para nosotros, dado que se realizó con un tamaño muestral similar al presente estudio (50 informes médicos en español o catalán de pacientes afectados por una enfermedad rara). Los resultados mostraron una sorprendente variabilidad, desde *bastante fáciles* hasta *muy difíciles*, lo cual es notable si se tiene en cuenta que todos los documentos pertenecían al mismo género textual y tenían una longitud similar. En contraste, nuestro análisis arroja una variación mucho menor, atribuible al hecho de que todos los informes provienen de un único servicio hospitalario, mientras que el estudio referido abarcaba informes de ocho especialidades diferentes.

En cuanto a la comprensibilidad de textos específicamente dirigidos a pacientes con patologías vasculares, las investigaciones son limitadas. Y aún son más escasas en español. Por eso, resulta relevante el análisis de San Norberto et al. (2024), aunque no se centre en informes médicos, sino en la calidad de la información disponible en internet sobre el aneurisma de aorta y su tratamiento endovascular. Este análisis de 180 páginas web en español reveló que, en general, el contenido en la web es deficiente en términos de accesibilidad, utilidad y confiabilidad, y clasificó su legibilidad como *algo difícil*. Observamos que estas páginas web accesibles para el público general, aunque todavía presentan dificultades en la comprensión debido a los términos utilizados, obtienen puntuaciones más altas en los índices de legibilidad en comparación con los informes de alta que hemos estudiado. Esto podría explicarse por el hecho de que las páginas web están diseñadas desde el principio para ser entendidas por personas sin conocimientos específicos sobre medicina, quienes las visitan en busca de respuestas a sus dudas médicas.

Si bien los resultados son alentadores, su aplicabilidad a otros entornos o herramientas de IA no es automática. Cada modelo puede presentar sesgos o estrategias diferentes para simplificar la información médica, lo que influye directamente en la comprensión del texto por parte del paciente. Evaluar de forma comparativa el desempeño de distintos modelos permitirá determinar con mayor precisión el alcance real de la mejora en legibilidad observada.

La principal limitación de nuestro estudio es que los índices de legibilidad no están diseñados específicamente para textos médicos especializados. Las fórmulas aplicadas se centran en la longitud de las palabras y las frases, pero no consideran si los términos están circunscritos al ámbito médico, como en el caso de los epónimos de enfermedades o nombres de medicamentos. Asimismo, no distinguen si una palabra breve forma parte de la jerga médica o si su significado es claro solo para quienes están familiarizados con las siglas pertinentes. No obstante, pese a estas limitaciones, siguen proporcionando una evaluación rápida y objetiva, lo que los convierte en herramientas valiosas para un análisis preliminar de la comprensibilidad de los textos médicos y explica su uso continuado en el ámbito sanitario. Además, para compensar este inconveniente, al menos en parte, se realizó un análisis de la precisión terminológica y semántica, de forma específica e independiente.

Todos estos estudios coinciden en subrayar la necesidad de adaptar el lenguaje escrito médico para hacerlo más accesible a todos los pacientes, poniendo de manifiesto que la claridad y la facilidad son fundamentales y deben ser consideradas estándares imprescindibles en la comunicación médica.

Hay investigaciones previas que exploran el uso de nuevas tecnologías para perfeccionar la comunicación sanitaria escrita, como la llevada a cabo por Estopà, et al. (2020) y que sugirieron la posibilidad de *enriquecer* los informes médicos de manera automática, para mejorar su claridad y comprensión, transformando términos técnicos y acrónimos. Sin embargo, nuestro proyecto se distingue por implementar una herramienta de la IA, un chatbot. En este caso, ChatGPT no solo simplifica el texto, sino que personaliza el contenido al seleccionar y reorganizar la información crucial para el paciente, además de omitir detalles que podrían herir su sensibilidad o afectar la relación médico-paciente.

La reescritura de informes médicos con la IA para adaptarlos al paciente es una vía prometedora que todavía se encuentra en una fase exploratoria.

revistadelta
41.3
2025

Aunque este estudio es una primera aproximación que sugiere su viabilidad, es esencial proseguir con la investigación. Su efectividad debe ser evaluada mediante estudios adicionales que examinen la comprensión de un grupo representativo de pacientes, para confirmar que la tecnología mejora la claridad y la accesibilidad de la información médica sin sacrificar la calidad. Uno de los principales beneficios de esta tecnología radica en su capacidad para procesar información de manera automática y muy rápida —menos de 30 segundos por informe en la actualidad— centrando la atención en el paciente y sin obstaculizar la transmisión de la información médica especializada para otro colega.

No obstante, como ya ha sido destacado por Sallam (2023), una de las principales limitaciones del uso de la IA para la reedición de textos es la generación de información falsa. En su revisión de 60 artículos editados con ChatGPT, encontraron información defectuosa en el 96,7% de los casos. Esto incluía "cuestiones éticas, derechos de autor y transparencia, cuestiones legales, riesgo de sesgo, plagio, falta de originalidad, citas incorrectas y tendencia a generar alucinaciones", entre otras.

Además, Abdelhafiz et al. (2024) resaltaron que el fenómeno de las alucinaciones es mayor en modelos entrenados sin supervisión externa, destacando la importancia de la evaluación humana para garantizar la precisión y validez del contenido generado.

Los resultados de nuestro estudio sugieren que la mayoría de los errores en la reedición de informes clínicos surgen de una reformulación incorrecta de la terminología (2,1‰), ya sea por imprecisión u omisión de datos. Sin embargo, a diferencia de lo observado en los estudios previamente mencionados, hemos notado que la interpretación errónea y la generación de alucinaciones por parte del ChatGPT no son frecuentes (0,18‰). Aunque la tasa de errores es muy baja, esta es una de sus principales limitaciones, por lo que parece aconsejable que el informe final sea revisado por profesionales para asegurar su precisión.

Respecto a la relación coste-beneficio de la implementación de chatbots, parece ser favorable, especialmente al evaluar el potencial ahorro de tiempo: los pacientes, mejor informados gracias a las explicaciones claras y detalladas en los informes, podrán formular preguntas más específicas y pertinentes en sus interacciones con el personal de la salud, lo que a su vez puede conducir a una atención más eficiente. Por otra parte, es recomendable que, al integrar estos sistemas en la infraestructura sanitaria existente, se

adopten medidas para garantizar la anonimización de los datos del paciente, evitando que la información personal identificable sea procesada por la IA.

En definitiva, la incorporación de chatbots a los sistemas de atención sanitaria tiene el potencial de mejorar la calidad del cuidado del paciente. Su uso podría mejorar la toma de decisiones compartidas y facilitar el empoderamiento del paciente, proporcionando informaciones de calidad y relevancia para cada caso específico.

5. Conclusiones

Este estudio demuestra que la inteligencia artificial, específicamente el uso de ChatGPT, puede mejorar significativamente la comprensibilidad de los informes médicos para los pacientes. A través de la aplicación de índices de legibilidad, se ha evidenciado que los informes adaptados por ChatGPT presentan una mejor legibilidad en comparación con los originales, logrando que estos sean más accesibles para un lector no especializado. Esta mejora es particularmente relevante en el contexto sanitario, donde la claridad en la comunicación es esencial para fomentar una relación de confianza entre el paciente y el profesional de la salud, así como para asegurar la comprensión adecuada de la información médica.

No obstante, aunque los resultados son prometedores, es importante considerar las limitaciones de la herramienta de inteligencia artificial utilizada. A pesar de la mejora en la legibilidad, se detectaron errores en la precisión médica, principalmente relacionados con la terminología, la medicación, las abreviaturas y algunas alucinaciones generadas por el chatbot. Aunque la tasa de errores fue baja, estas inexactitudes podrían comprometer la calidad y la seguridad de la información proporcionada a los pacientes. Por ello, se recomienda una revisión cuidadosa y supervisión por parte de profesionales médicos antes de utilizar informes generados por IA de manera autónoma.

Entre las limitaciones del presente estudio se encuentra la dependencia de un único modelo de IA (ChatGPT-3.5). Las variaciones entre plataformas de inteligencia artificial pueden influir en los resultados obtenidos. Futuros estudios deberían explorar si las mejoras de legibilidad y las tasas de error se mantienen al utilizar otros modelos o versiones más avanzadas.

revistadelta
41.3
2025

En resumen, el uso de la inteligencia artificial para adaptar informes médicos presenta una valiosa oportunidad para mejorar la comunicación escrita en el ámbito sanitario. Sin embargo, su implementación debe realizarse con precaución, asegurando siempre la intervención de expertos para validar la precisión y la pertinencia de la información adaptada. De cara al futuro, sería conveniente seguir investigando y desarrollando estas herramientas, con el fin de minimizar los errores y optimizar su uso para el beneficio tanto de los pacientes como de los profesionales de la salud.

Fuentes de financiación

Este trabajo ha sido financiado por las Ayudas a la Investigación de la Cátedra de Humanización de la Asistencia Sanitaria de VIU, Fundación ASISA y Proyecto HUCI (nº de referencia 24I288.01/1).

Declaración de conflictos de interés

Los autores de este artículo declaran no tener conflictos de intereses financieros, profesionales o personales que pudieran haber influido de manera inapropiada en este trabajo.

Declaración de contribución de autoría

Adéla Koťátková y Manuel Miralles Hernández concibieron y supervisaron el estudio, participaron en la conceptualización, en la elaboración de la metodología y en el análisis formal. Irene Peinado Zanón, María Álamo Rodríguez, Alejandra López Hernández y Emma Plana participaron como investigadoras, realizaron el análisis formal, la metodología y las revisiones del manuscrito. Todos los autores participaron en la redacción, revisión y edición del manuscrito, aprobaron la versión final y se responsabilizan de todos los aspectos del trabajo, garantizando la veracidad y la integridad de su contenido.

Disponibilidad de datos

Los datos relevantes que sustentan los resultados de este estudio están incluidos en el artículo.

Referencias

Abdelhafiz, A. S., Ali, A., Maaly, A. M., Ziady, H. H., Sultan, E. A., & Mahgoub, M. A. (2024). Knowledge, perceptions and attitude of researchers towards using ChatGPT in research. *Journal of Medical Systems*, *481*(26). https://doi.org/10.1007/s10916-024-02044-4.

Barrio Cantalejo, I. M., & Simón Lorda, P. (2003). ¿Pueden leer los pacientes lo que pretendemos que lean? Un análisis de la legibilidad de materiales escritos de educación para la salud. *Atención Primaria*, *31*(7), 409-414. https://doi.org/10.1016/s0212-6567(03)79199-9.

Barrio-Cantalejo, I. M., Simón-Lorda, P., Melguizo, M., Escalona, I., Marijuán, M. I., & Hernando, P. (2008). Validación de la escala INFLESZ para evaluar la legibilidad de textos dirigidos al paciente. *Anales del Sistema Sanitario de Navarra*, *31*(2), 135-152. https://doi.org/10.4321/s1137-66272008000300004.

Blanco Pérez, A., & Gutiérrez Couto, U. (2002). Legibilidad de las páginas web de salud para pacientes y lectores de la población general. *Revista Española de Salud Pública*, *76*(4), 321-331.

Estopà, R., & Domènech-Bagaria, O. (2019). Diagnóstico del nivel de comprensión de informes médicos dirigidos a pacientes y familias afectados por una enfermedad rara. *E-Aesla*, *5*, 109-118.

Estopà, R., López-Fuentes, A., & Porras-Garzón, J. M. (2020). Terminology in written medical reports: A proposal of text enrichment to favour its comprehension by the patient. En B. Daille, K. Kageura, & Rigouts Terryn, A. (Eds), *Proceedings of the 6th International Workshop on Computational Terminology* (pp. 43-49). European Language Resources Association.

Fernández Huerta, J. (1959). Medidas sencillas de lecturabilidad. *Consigna*, *2*(14), 29-32.

Gómez-González, E., Priego Álvarez, H. R., & García, C. de la C. (2023). Empoderamiento y adherencia terapéutica en el adulto mayor: una revisión sistemática. *Enfoque*, *33*(29), 46-63.

Muñoz Fernández, A. (2018). *Analizador de legibilidad de texto*. https://legible.es/.

Muñoz, B., y Muñoz, U. (2019). *Legibilidad Mu* [Disponible en http://www.legibilidadmu.cl]. Viña del Mar, Chile.

Pelicano-Piris, N. (2022). Análisis de la legibilidad de la información sanitaria dirigida a la población vulnerable. *Revista ProPulsión*, *5*(2), 140-171. https://doi.org/10.53645/revprop.v5i2.95.

Porras-Garzón, J. M., y Estopà, R. (2020). Escalas de legibilidad aplicadas a informes médicos: límites de un análisis cuantitativo formal. *Círculo de Lingüística Aplicada a la Comunicación*, *83*, 205-216. https://doi.org/10.5209/clac.70574.

revistadelta
41.3
2025

Ramírez-Puerta, M. R., Fernández-Fernández, R., Frías-Pareja, J. C., Yuste-Ossorio, M. E., Narbona-Galdo, S., y Peñas-Maldonado, L. (2013). Análisis de la legibilidad del consentimiento informado en cuidados intensivos. *Medicina Intensiva*, *37*(8), 503-509. https://doi.org/10.1016/j.medin.2012.08.013.

Rubiera, G., Arbizu, R., Alzueta, A., Agúndez, J. J., & Riera, J. R. (2004). Legibilidad de los documentos de consentimiento informado utilizados en los hospitales de Asturias. *Gaceta Sanitaria*, *18*(2), 153-158. https://doi.org/10.1016/s0213-9111(04)71822-1.

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, *11*(6), 887. https://doi.org/10.3390/healthcare11060887.

San Norberto, E. M., Revilla, A., Brizuela, J. A., & Taylor, J. H. (2024). Quality and readability of Spanish-language online information for aortic aneurysm and its endovascular treatment. *Vascular and Endovascular Surgery*, *58*(2), 158-165. https://doi.org/10.1177/15385744231196644.

Simón Lorda, P., Barrio Cantalejo, I. M., y Concheiro Carro, L. (1996). Legibilidad de los formularios escritos de consentimiento informado. *Medicina Clínica*, *107*(14), 524-529.

Szigriszt Pazos, F. (1993). *Sistemas predictivos de legibilidad del mensaje escrito: fórmula de perspicuidad* (tesis doctoral). Universidad Complutense de Madrid.

Recebido em: 16.09.2024
Aprovado em: 19.10.2025