Evaluación de la precisión, claridad, relevancia y legibilidad de ChatGPT 4.0 en respuestas a preguntas frecuentes de pacientes sobre infertilidad

ChatGPT 4.0: accurate, clear, relevant, and readable responses to frequently asked fertility patient questions

Melina Schapira[1], Micaela Montiveros[1], Fiamma Di Biase[1], Carolina Formica Muntaner[1], Sergio Papier[1], Demian Glujovsky[1].

[1] Cegyr

RESUMEN

Pregunta de estudio: ¿Las respuestas generadas por ChatGPT 4.0 a preguntas frecuentes de pacientes con infertilidad presentan adecuada precisión, claridad, relevancia y legibilidad?
Respuesta resumida: ChatGPT 4.0 brindó respuestas de buena calidad global, con fortalezas en apoyo emocional, pronóstico y estilo de vida, y legibilidad adecuada.
Lo que ya se sabe: Los pacientes con infertilidad recurren cada vez más a la inteligencia artificial para obtener información, pero existe escasa evidencia en español y en medicina reproductiva sobre la calidad de estas respuestas.
Diseño del estudio: Estudio transversal observacional. Se analizaron 50 preguntas frecuentes; el estudio se desarrolló durante 2024.
Materiales y Métodos: Se seleccionaron 50 preguntas recopiladas de foros y blogs de pacientes con infertilidad. Las respuestas fueron generadas por ChatGPT 4.0 mediante un prompt estandarizado solicitando responder como especialista en infertilidad. Diez especialistas (5 senior y

ABSTRACT

Objective: *To evaluate the accuracy, clarity, relevance and readability of ChatGPT 4.0's responses to frequently asked infertility questions.*
Materials and Methods: *Cross-sectional study. Fifty frequently asked questions were collected from online patient forums. ChatGPT 4.0 generated responses using a standardized prompt ("as if ChatGPT were an infertility specialist, using the best available evidence"). Ten reproductive medicine specialists (5 senior, 5 junior) evaluated answers for accuracy, clarity, and relevance using Likert scales (1–5). Overall quality was assessed with the Global Quality Scale (GQS). Readability was determined with the Spanish Flesch-Kincaid index.*
Results: *94% of responses scored ≥3 in GQS (mean 3.6 ± 0.6). 62% were rated "very good" or "excellent". Highest ratings were observed in emotional support (4.4–4.5), prognosis (4.2) and lifestyle (4.2–4.3). Infertility diagnosis showed lower accuracy (3.7). All specialists (100%) agreed ChatGPT could be a useful*

ENVÍO CORRESPONDENCIA: xxxxxx

5 junior) evaluaron precisión, claridad y relevancia mediante escalas Likert (1–5). La calidad global se midió con la Global Quality Scale (GQS). La legibilidad se evaluó mediante el índice Flesch-Kincaid en español. El estudio fue exento de evaluación ética al no involucrar participantes humanos ni datos sensibles.

Resultados: El 94% de las respuestas obtuvo un puntaje ≥3 en la GQS (media 3,6 ± 0,6). El 62% fue calificada como "muy buena" o "excelente". Los puntajes más altos correspondieron a apoyo emocional (4,4–4,5), pronóstico (4,2) y estilo de vida (4,2–4,3). El menor desempeño fue diagnóstico de infertilidad (3,7). Todos los especialistas consideraron que ChatGPT podría utilizarse como herramienta complementaria bajo supervisión médica. La legibilidad media fue 19,6 ± 4,2.

Limitaciones del estudio: Incluye un número limitado de preguntas, evalúa una sola versión del modelo de IA y no compara con otras herramientas o materiales educativos.

Implicancias de los hallazgos: ChatGPT 4.0 podría utilizarse para mejorar la comprensión inicial del paciente, brindar apoyo entre consultas y reducir mitos o desinformación, siempre dentro de un marco supervisado por profesionales.

Palabras clave: Infertilidad; Inteligencia Artificial; Comunicación médico-paciente; Consejería; ChatGPT.

complementary tool for patients, when supervised, and 7 of 10 rated its overall performance as "very good" or "excellent". Readability was 19.6 ± 4.2, equivalent to secondary school level.

Conclusion: *ChatGPT 4.0 provides clear and relevant responses, with acceptable accuracy, making it a potential complementary tool for initial infertility counseling. It does not replace medical consultation and requires professional oversight.*

Keywords: *infertility, artificial intelligence, ChatGPT, patient communication, counseling.*

INTRODUCCIÓN

La infertilidad afecta entre el 10 y el 15% de las parejas en edad reproductiva a nivel global[1]. Este diagnóstico suele generar una importante carga emocional relacionada con la incertidumbre, los tiempos de espera y la complejidad de los tratamientos de reproducción asistida. En este contexto, el acceso a información confiable y comprensible es un componente clave para el acompañamiento integral de las personas que atraviesan dificultades reproductivas.

En las últimas décadas, internet se consolidó como una de las principales fuentes de búsqueda de información en salud. La mayoría de los pacientes consulta foros, blogs, redes sociales y contenido biomédico de diversa calidad antes de acudir a un especialista. Aunque este fenómeno facilita el acceso a contenidos, también incrementa la exposición a datos incompletos, contradictorios o erróneos, lo que puede aumentar la ansiedad, generar falsas expectativas o incluso favorecer decisiones inapropiadas sobre la salud[2,3].

En paralelo, el desarrollo de herramientas basadas en inteligencia artificial (IA), especialmente los modelos de lenguaje como ChatGPT, introdujo una nueva forma de interacción para la obtención de información médica. Su capacidad para generar respuestas comprensibles y personalizadas lo vuelve atractivo para pacientes que buscan orientación inicial. Sin embargo, la información generada por IA no siempre es precisa, está influenciada por la calidad del prompt, está basada en información no corroborable, y carece de contexto clínico individual. En áreas altamente especializadas, como la reproducción asistida, no existe suficiente evidencia que respalde su uso seguro[5].

Una característica adicional es la falta de estudios en español y en población latinoamericana. Considerando las diferencias culturales, semánticas y lingüísticas, es fundamental evaluar cómo se desempeñan estas herramientas en este entorno específico.

El objetivo de este estudio fue analizar de manera sistemática la precisión, claridad, relevancia y legibilidad de las respuestas de ChatGPT 4.0 sobre temas de infertilidad. También se buscó explorar la percepción de especialistas respecto del potencial de esta tecnología como herramienta complementaria de información para pacientes.

MATERIALES Y MÉTODOS

Diseño del estudio

Se realizó un estudio de corte transversal observacional que evaluó la calidad de las respuestas generadas por ChatGPT 4.0 frente a preguntas frecuentes de pacientes sobre infertilidad.

Selección de preguntas

Se recopilaron 50 preguntas frecuentes provenientes de blogs, foros y espacios digitales utilizados habitualmente por pacientes que consultan sobre fertilidad. Se priorizaron preguntas representativas de temáticas recurrentes, tales como diagnóstico de infertilidad, tratamientos (IVF, ICSI, ovodonación), medicación y hormonas, pronóstico, estilo de vida, tiempos y plazos, apoyo emocional, tecnologías de reproducción asistida.

Las preguntas fueron seleccionadas por dos investigadores con experiencia en medicina reproductiva.

Generación de respuestas con ChatGPT

Cada pregunta fue ingresada en la versión ChatGPT 4.0 utilizando un prompt estandarizado: **“Respondé como si fueras un especialista en infertilidad, utilizando la**

mejor evidencia disponible."

Se registraron las respuestas completas sin edición.

Evaluación por especialistas

Un panel compuesto por 10 especialistas en medicina reproductiva evaluó cada respuesta. La mitad del panel tenía menos de 10 años de experiencia (junior) y la otra mitad más de 10 años (senior).

Herramientas de evaluación

- **Global Quality Scale (GQS) (1-5):** valoración de calidad global.
- **Escalas Likert (1–5):** precisión, claridad y relevancia.
- **Índice Flesch-Kincaid (español):** para determinar legibilidad.

Análisis estadístico

Se calcularon medias y desvíos estándar. Se realizaron comparaciones entre evaluadores senior y junior. Se consideró que la diferencia era estadísticamente significativa con un valor de $p < 0,05$, utilizando t-test para medidas continuas y chi cuadrado para proporciones Se utilizó STATA 17.0 para el análisis.

Aprobación ética

El estudio fue exento de evaluación por parte del Comité de Ética, al no involucrar participantes humanos ni datos clínicos sensibles.

RESULTADOS

Evaluación global

El 94% de las respuestas (47/50) obtuvo una puntuación ≥3 en la GQS. El promedio general fue 3,6 ± 0,6. El 62% fue clasificada como "muy buena" o "excelente".

Todos los especialistas (10/10) consideraron que ChatGPT puede ser una herramienta complementaria útil para pacientes que buscan información sobre infertilidad, siempre bajo supervisión médica. Siete de los diez evaluadores calificaron el desempeño global como "muy bueno" o "excelente". La distribución se encuentra graficada en la figura 1.

Resultados por dominios

Los resultados por dominios se presentan en la Tabla 1.

Legibilidad

El índice Flesch-Kincaid fue 19,6 ± 4,2,

Figura 1. Distribución de las valoraciones globales otorgadas por especialistas a las respuestas de ChatGPT 4.0. Se observa que el 50% calificó el desempeño como "muy bueno", el 30% como "excelente" y el 20% como "bueno".

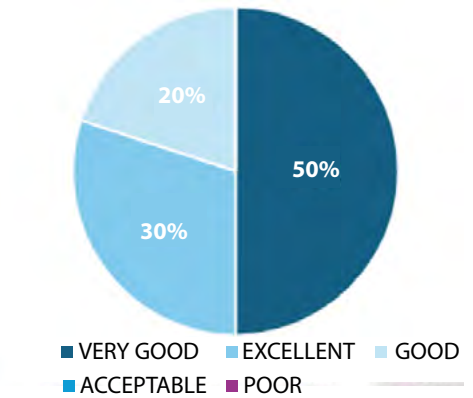

Tabla 1. Puntuaciones medias por dominio

Dominio	Precisión	Claridad	Relevancia
Diagnóstico de infertilidad	3,7	3,9	3,9
Tratamientos	4,0	4,1	4,2
Medicación	4,0	4,0	4,0
Pronóstico	4,2	4,2	4,2
Estilo de vida	4,2	4,2	4,3
Tiempos y plazos	3,9	4,1	4,2
Apoyo emocional	4,4	4,4	4,5
Tecnologías de RA	4,0	4,2	4,3

El dominio con menor desempeño fue diagnóstico de infertilidad (3,7 sobre 5), mientras que apoyo emocional obtuvo los puntajes más altos (4,5 sobre 5).

equivalente a nivel de lectura de 10° grado en sistema anglosajón (secundario superior), considerado "ligeramente difícil pero accesible".

DISCUSIÓN

Los resultados muestran que ChatGPT 4.0 genera respuestas de buena calidad en las preguntas de la mayoría de los dominios de infertilidad evaluados. La claridad y relevancia fueron particularmente destacadas, lo que coincide con observaciones previas sobre la capacidad de los modelos de IA para sintetizar conceptos complejos en lenguaje accesible[5].

El desempeño más bajo se observó en el dominio diagnóstico. Esto es esperable, dado que el diagnóstico en medicina reproductiva requiere integrar información clínica específica, estudios complementarios y juicios profesionales que no pueden extrapolarse a partir de una pregunta aislada. La IA puede ofrecer definiciones, criterios diagnósticos y explicaciones generales, pero no puede identificar causas específicas sin información personalizada.

La buena valoración en "apoyo emocional" resulta especialmente relevante. Los pacientes suelen buscar contención, claridad y validación de sus preocupaciones. ChatGPT mostró fortalezas para ofrecer explicaciones empáticas y estructuradas, lo cual podría ser útil como herramienta previa a la consulta, reduciendo ansiedad y mejorando la comprensión inicial.

Potenciales aplicaciones clínicas

La utilización de ChatGPT podría utilizarse en la confección de material educativo preliminar, reduciendo probablemente el tiempo y facilitando la producción de dicho material, con el control y edición necesario de profesionales médicos. Si bien no fue probado para esos fines, podría evaluarse la utilización de confección de material de apoyo entre consultas, ayudando a una mejor comunicación con el paciente. Asimismo, podría facilitar la preparación para la primera consulta mediante la entrega de material de lectura, y constituir una herramienta complementaria para disminuir mitos o desinformación.

Limitaciones

El uso independiente por parte del paciente no está recomendado puesto que depende del diseño del prompt y la valoración de un profesional sobre el contenido. Aunque es mayoritariamente confiable en función de los resultados encontrados, no

se puede concluir que la utilización sin la supervisión profesional sea confiable. Las respuestas que entrega son genéricas, por lo que la utilización en determinado contexto clínico no fue evaluado, perdiendo así la personalización de la respuesta. Cabe recalcar que se utilizó ChatGPT para responder preguntas frecuentes de blogs o foros, y no para responder preguntas en consulta de pacientes, por lo que no se pueden extrapolar estos resultados a la consulta clínica. Se evaluó con ChatGPT 4.0 por lo que la extrapolación a otras IA es inadecuada. No se evaluaron temas controversiales que pudieran dar lugar a valoraciones éticas.

Perspectivas futuras

Será importante evaluar nuevas versiones de IA, su desempeño en español y su impacto real en pacientes, incluyendo métricas como ansiedad, satisfacción y adherencia a tratamientos.

CONCLUSIONES

ChatGPT 4.0 mostró un desempeño consistentemente bueno en la mayoría de los dominios analizados para preguntas realizadas en foros o blogs, con especial fortaleza en apoyo emocional, estilo de vida y pronóstico. La precisión fue aceptable y la claridad elevada. La totalidad de los especialistas coincidió en su potencial como herramienta complementaria.

No debe reemplazar la consulta médica, pero sí puede ocupar un rol valioso como recurso de orientación inicial para pacientes que se enfrentan por primera vez a dudas sobre infertilidad.

REFERENCIAS

1. World Health Organization. Infertility prevalence estimates, 1990–2021. Geneva: WHO; 2023.
2. Inhorn MC, Patrizio P. Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century. Hum Reprod Update. 2015;21(4):411–26.
3. Pedro J, et al. Patients' attitudes towards the use of artificial intelligence in reproductive medicine. Hum Reprod. 2022;37(11):2602–12.
4. Cascella M, et al. Evaluating the accuracy of ChatGPT in medical information: systematic review. J Med Internet Res. 2023;25:e48602.