Rev Argent Cir. 2026;118(2):e-1963- http://dx.doi.org/10.25132/raac.v118.n2.1963

Exactitud de la inteligencia artificial en la estadificación y toma de decisiones en cáncer esófago gástrico

Accuracy of artificial intelligence for staging and decision-making in esophageal and gastric cancer

Adelina E. Coturel, Paula Pereyra, Rodrigo García, Agustín Diomedi, Juan J. M. Cabas Audicio, Roberto Klappenbach

Institución: Hospital de Alta Complejidad del Bicentenario de Esteban Echeverria. Buenos Aires, Argentina.

Los autores declaran no tener conflictos de interés.
Conflicts of interest None declared.

Correspondencia *Correspondence:* Adelina E. Coturel
E-mail: adelinacoturel@gmail.com

RESUMEN

Antecedentes: la inteligencia artificial (IA) ha mostrado un creciente potencial en el ámbito de la oncología, no solo en el análisis de imágenes sino también como herramienta de apoyo en la toma de decisiones clínicas.
Objetivo: evaluar un modelo conversacional basado en IA, entrenado con guías clínicas, para asistir en la estadificación y planificación terapéutica inicial en cáncer esofagogástrico.
Material y métodos: se realizó un estudio retrospectivo en pacientes con cáncer de esófago, estómago y de la unión esofagogástrica, en un hospital de tercer nivel de atención, tratados entre diciembre de 2023 y mayo de 2025. Se desarrolló un asistente virtual o chatbot con ChatGPT-4.5® (OpenAI), configurado para interpretar los estudios según guías NCCN y ESMO. El chatbot debía determinar el estadio clínico (cTNM), sugerir la conducta terapéutica y detectar hallazgos relevantes. Sus respuestas se compararon con las decisiones del equipo médico tratante.
Resultados: de 53 pacientes evaluados, 34 cumplieron los criterios de inclusión. La concordancia entre el cTNM y el estadio clínico asignados por el equipo médico y el chatbot fue del 94%. La coincidencia en la conducta terapéutica sugerida fue del 85%. Las discrepancias fueron por diferencias en la interpretación del estadio, consideraciones clínicas no disponibles para la IA y aspectos sociales o logísticos que influyeron en la decisión médica.
Conclusión: la IA aplicada mediante un chatbot mostró utilidad para asistir en la estadificación y propuesta terapéutica inicial del cáncer esofagogástrico, con alta concordancia con las decisiones clínicas.

Palabras clave: *inteligencia artificial, neoplasias esofágicas, neoplasias gástricas, estadificación de tumores, sistemas de apoyo a la toma de decisiones clínicas.*

ABSTRACT

Background: Artificial intelligence (AI) has demonstrated increasing potential in oncology, including in imaging analysis and as a tool for supporting clinical decision-making.
Objective: The aim of this study was to evaluate an AI conversational model, trained using up-to-date clinical guidelines, to assist with staging and initial treatment planning in patients with gastroesophageal cancer.
Materials and methods: We conducted a retrospective study of patients with esophageal, gastric, and gastroesophageal junction cancer treated at a tertiary care hospital between December 2023 and May 2025. A virtual clinical assistant or chatbot was developed using ChatGPT-4.5® (OpenAI), configured to interpret test results in accordance with the NCCN and ESMO guidelines. The chatbot was designed to determine the clinical stage (cTNM), suggest a treatment plan, and identify relevant findings. The answers were compared with the decisions made by the treating physician.
Results: Of the 53 patients evaluated, 34 met the inclusion criteria. The accuracy between the cTNM staging determined by the medical team and that of the chatbot was 94%. The concordance in the suggested therapeutic approach was 85%. The discrepancies were due to differences in the interpretation of the stage, clinical considerations not available to the AI, and social or logistical factors that influenced the medical decision.
Conclusion: AI, implemented through a chatbot trained using clinical guidelines, proved useful in supporting staging and initial treatment planning for gastroesophageal cancer, with high concordance with clinical decisions.

Keywords: *artificial intelligence, esophageal neoplasms, stomach neoplasms, neoplasm staging, decision support systems, clinical.*

Recibido | *Received* 01-01-26
Aceptado | *Accepted* 01-04-26

ID ORCID: Adelina E. Coturel, 0000-0001-5379-6872; Paula Pereyra, 0009-0007-2682-586X; Rodrigo García, 0000-0001-9442-1989; Agustín Diomedi, 0009-0009-0961-2461; Juan J. M. Cabas Audicio, 0009-0007-4379-6941; Roberto Klappenbach, 0000-0002-0069-0035.

Introducción

La inteligencia artificial (IA) se ha consolidado como una herramienta transformadora en medicina, especialmente en oncología, donde ha demostrado mejorar la precisión diagnóstica, optimizar la planificación terapéutica y apoyar la toma de decisiones personalizadas[1]. En términos generales, la IA comprende sistemas capaces de reproducir funciones cognitivas humanas, como el aprendizaje y el razonamiento[2]. Dentro de este campo, los modelos de lenguaje natural utilizados en asistentes conversacionales, *chatbots*, permiten interpretar e interactuar con lenguaje clínico complejo.

El avance del *deep learning* ha ampliado la capacidad de estas tecnologías para analizar información heterogénea, imágenes, texto libre y datos clínicos estructurados, y vincularla con guías de práctica clínica. En oncología digestiva se han informado aplicaciones exitosas en detección temprana de cáncer esofagogástrico, identificación automatizada de lesiones endoscópicas y predicción de respuesta terapéutica[3]. Sin embargo, hay poca evidencia sobre el uso de modelos conversacionales para el desarrollo de asistentes que permitan colaborar con la estadificación y planificación inicial del tratamiento.

En la Argentina, la evidencia sobre el uso de la IA en medicina aún es limitada. Se han publicado experiencias en reconocimiento de imágenes para la visión crítica de seguridad en colecistectomías en el Hospital Argerich[4], en la evaluación del uso de ChatGPT para recomendaciones de *screening* de cáncer colorrectal[5], y en la implementación de un copiloto de IA integrado a historias clínicas electrónicas en el Hospital Italiano de Buenos Aires[6], orientado a mejorar el acceso a información relevante del paciente.

En este contexto, el objetivo principal de nuestro estudio fue evaluar la precisión de un modelo conversacional de IA para predecir el estadio clínico (cTNM y estadio global) a partir de los estudios disponibles en la primera consulta oncológica de pacientes con cáncer esofagogástrico. Como objetivo secundario, analizar la concordancia entre las conductas terapéuticas sugeridas por el *chatbot* y las indicadas por el comité de tumores o el equipo tratante.

Material y métodos

Se realizó un estudio retrospectivo sobre la base de datos de pacientes que consultaron por primera vez, por patología oncológica, en el consultorio de cirugía esofagogástrica de la institución, entre diciembre de 2023 y mayo de 2025.

Fueron incluidos aquellos pacientes de los cuales se contaba con estudios preoperatorios completos, definidos como: pacientes con diagnóstico de tumor de esófago, unión esofagogástrica (UEG) o estómago, informe de endoscopia digestiva alta, informe anatomopatológico de biopsia, tomografía computarizada, tomografía por emisión de positrones (PET-TC) en pacientes con cáncer de esófago o tumores de la UEG.

Fueron excluidos los pacientes sin la totalidad de los estudios preoperatorios mencionados, o con histología diferente de adenocarcinoma o carcinoma epidermoide.

Las decisiones terapéuticas se tomaron en forma multidisciplinaria entre cirujanos especialistas en Cirugía Esofagogástrica oncológica, oncólogos clínicos, anestesiólogos, médicos clínicos y nutricionistas.

Desarrollo y uso del asistente virtual

Para el análisis automatizado se desarrolló un asistente clínico virtual utilizando ChatGPT (OpenAI), versión GPT-4.5®, a través de una suscripción paga ChatGPT Plus®. El desarrollo y configuración del asistente fue realizado por cirujanos con formación en Cirugía Esofagogástrica y experiencia en estadificación tumoral. Si bien no cuentan con formación formal en programación ni en desarrollo algorítmico, poseen capacitación en IA aplicada al ámbito sanitario, lo que permitió estructurar el uso del modelo bajo criterios clínicos y metodológicos adecuados.

El *chatbot* fue configurado específicamente para esta tarea mediante la carga y lectura contextual de las guías clínicas actualizadas de la Red Nacional del Cáncer de Estados Unidos (NCCN, por sus siglas del inglés National Comprehensive Cancer Network)[7,8] y de la Sociedad Europea de Oncología Clínica (ESMO, del inglés European Society for Medical Oncology)[9,10], correspondientes a cáncer de esófago y cáncer gástrico.

El modelo fue instruido para procesar, en lenguaje natural, la información clínica de cada paciente simulando una consulta oncológica inicial. Cabe destacar que, para ello, no fue reentrenado ni modificado a nivel algorítmico. Se trabajó mediante ingeniería de *prompts*, definiendo instrucciones estructuradas basadas en las guías internacionales mencionadas. A partir de los datos aportados (edad, sexo, comorbilidades, motivo de consulta, informes de endoscopia, anatomía patológica e imágenes), el asistente debía: a) determinar el estadio clínico del tumor según la clasificación TNM 8ª edición (AJCC)[11]; b) sugerir el tratamiento oncológico inicial adecuado según guías internacionales.

Previamente al análisis definitivo, se realizaron pruebas piloto con casos históricos no incluidos en la cohorte final, con el objetivo de estandarizar el formato de ingreso de datos, ajustar la redacción de las instrucciones y homogeneizar la estructura de las respuestas. No se realizaron modificaciones sustanciales en las instrucciones durante el análisis de los casos incluidos en el estudio.

Cada caso de estudio fue procesado de manera

individual, cargando los datos clínicos en la plataforma y registrando las respuestas generadas por la IA en una tabla estructurada. Posteriormente, las propuestas del asistente fueron comparadas con las decisiones tomadas por el comité de tumores o el equipo tratante en la práctica clínica habitual.

En aquellos casos en los que se identificaron discrepancias en la estadificación o en la conducta terapéutica sugerida, se realizó una interacción dirigida con el *chatbot*, con el objetivo de explorar las causas de la diferencia y determinar si se trataba de una interpretación alternativa válida, una omisión de datos relevantes, o una limitación en la comprensión contextual de los estudios.

Análisis estadístico

Las variables analizadas fueron:

1) demográficas: edad, sexo, comorbilidades, motivo de consulta; 2) del tumor: topografía, histopatología, informes de endoscopia, tomografía y PET-TC, cTNM, estadio clínico; 3) del tratamiento: realización de tratamiento endoscópico, tratamiento quirúrgico sin neoadyuvancia, tratamiento quirúrgico con neoadyuvancia o quimioterapia perioperatoria, tratamiento paliativo.

Se realizó un análisis descriptivo de los datos recolectados. Las variables categóricas se expresaron como frecuencias absolutas y porcentajes, mientras que las variables continuas se informaron como media y rango.

La concordancia entre la estadificación clínica (cTNM y estadio global) propuesta por el *chatbot* y la realizada por el equipo médico fue evaluada mediante comparación directa caso por caso. Del mismo modo, se analizó la concordancia en la conducta terapéutica sugerida.

El estudio se diseñó bajo la hipótesis de que un modelo conversacional de IA, entrenado con guías clínicas actualizadas, podría predecir correctamente la estadificación clínica en al menos el 80% de los casos y sugerir tratamientos acordes con las recomendaciones internacionales. Dado el carácter exploratorio del estudio y el tamaño de la muestra, no se aplicaron pruebas estadísticas inferenciales.

La exactitud del modelo se calculó como la proporción de casos en los que la respuesta del *chatbot* coincidió con la estadificación o la conducta propuesta por el equipo médico, sobre el total de casos evaluados. Se utilizó la siguiente fórmula:

$$\text{Exactitud (\%)} = (\text{Casos concordantes} / \text{Total de casos}) \times 100$$

Se calcularon por separado la exactitud para cTNM, estadio clínico y tratamiento sugerido.

Consideraciones éticas: El presente estudio fue realizado de acuerdo con los principios éticos de la Declaración de Helsinki. Por tratarse de un estudio retrospectivo, observacional y basado en el análisis de datos clínicos anonimizados, no se requirió consentimiento informado.

Resultados

De 53 pacientes que consultaron por tumores de esófago, estómago y de la UEG en el período indicado, se excluyeron 19 por falta de datos relevantes en la historia clínica, y se incluyeron 34 casos que cumplían los criterios de inclusión. De ellos, 18 (51%) eran de sexo masculino y la edad promedio fue de 67 años (rango 51-80). En 29 pacientes se encontraba detallado el motivo de consulta, y eso se analizó en la tabla 1, junto al resto de los datos demográficos y la ubicación del tumor.

TABLA 1

Variables demográficas y de los tumores

Variables	n (%)
Comorbilidades	
Diabetes mellitus	11 (32%)
HTA	19 (56%)
EPOC	6 (18%)
Tabaquista o extabaquista	9 (27%)
Enfermedad cardiovascular	9 (27%)
Cáncer de mama	6 (18%)
Otros cánceres	4 (11%)
Motivo de consulta	
Hemorragia digestiva	6 (18%)
Disfagia	16 (47%)
Dispepsia o dolor epigástrico	14 (41%)
Síndrome pilórico o vómitos	9 (26%)
Descenso de peso	26 (76%)
Ubicación del tumor	
Esófago	10 (29%)
Estómago	13 (38%)
Unión esofagogástrica	11 (32%)
- Siewert I	2 (18%)
- Siewert II	6 (54%)
- Siewert III	3 (27%)
Anatomía patológica	
Células en anillo de sello	
Esófago	1 (10%)
Unión esofagogástrica	3 (27%)
Estómago	4 (31%)
Carcinoma escamoso*	5 (50%)

*Del total de los cánceres de esófago.
HTA: hipertensión arterial; EPOC: enfermedad pulmonar obstructiva crónica.

Según el estadio, 3 pacientes presentaron estadio I, 7 estadio II, 17 estadio II y 7 estadio IV. El resto de las características se describen en la tabla 2.

La exactitud entre el cTNM estipulado por el equipo médico y el del *chatbot* fue del 94%. Dentro del tratamiento propuesto, la concordancia fue del 85%.

Los casos discordantes se presentan en la tabla 3.

Al dividir la cohorte en tres grupos secuenciales (n = 12, 11 y 11 pacientes), encontramos una mejoría en la precisión en el cTNM y el estadio del *chatbot*: 83% en el grupo 1, 100% en el grupo 2 y también 100% en el grupo 3.

■ TABLA 2

Comparación entre la estadificación clínica realizada por el equipo médico y el chatbot

cTNM**	Equipo médico	Chatbot
cT		
cT1b	1(3%)	1 (3%)
cT1b-2	1(3%)	1 (3%)
cT2-3	3(9%)	3 (9%)
cT3	21(62%)	21 (62%)
cT3-4	4(12%)	4 (12%)
cT4	4(12%)	4 (12%)
cN		
cN0-1	1(3%)	1 (3%)
cN+	20(59%)	21 (62%)
cM		
cM1	7(21%)	6(18 %)
Estadios		
I	3(9%)	3 (9%)
II	7(20%)	6 (18%)
III	17(50%)	19 (56%)
IV	7(20%)	6 (18%)

Discusión

El presente estudio evaluó la utilidad de un modelo conversacional de inteligencia artificial, entrenado con guías clínicas actualizadas, para asistir en la estadificación y planificación terapéutica inicial en pacientes con cáncer de esófago, unión esofagogástrica y estómago. Los resultados muestran una alta concordancia entre la estadificación clínica propuesta por el *chatbot* y la realizada por el equipo médico tratante (94%), tanto a nivel de cTNM como de estadio global. Si bien la concordancia en la conducta terapéutica fue algo menor (85%), en la mayoría de los casos las diferencias se explicaron por consideraciones clínicas contextuales no siempre disponibles para el modelo, como limitaciones sociales.

Los *chatbots* son programas que procesan y simulan conversaciones humanas mediante lenguaje natural[12]. Su uso en salud creció durante la pandemia de COVID-19, aunque aún permanece subutilizado y se espera que aumente en la próxima década. En cáncer gástrico, distintos modelos de inteligencia artificial han mostrado precisión prometedora para identificar individuos de alto riesgo, predecir agresividad tumoral, recurrencia y supervivencia[13,14]. El deep learning ha demostrado capacidad para analizar imágenes endoscópicas, datos genómicos e histopatológicos, mejorando diagnóstico y pronóstico. La integración multimodal de datos clínicos, genéticos e imagenológicos permite obtener una visión más completa de la enfermedad y

■ TABLA 3

Casos discordantes entre el Equipo Médico y el chatbot

Caso	Categoría de la discrepancia	Motivo	Equipo Médico	chatbot
1	cTNM y estadio clínico	Adenopatías lateroaórticas en estudios por imágenes	Considera M1, estadio IV	Considera adenopatías regionales, M0, estadio III
1	Tratamiento	Discrepancia de estadio clínico (IV vs. III)	Tratamiento paliativo	Neoadyuvancia y cirugía
2	cTNM y estadio clínico	Tumor de la UEG, sin adenopatías hipercaptantes en la PET. El chatBot lo considera N + probable por la longitud del tumor (5,8 cm) y SUV 13,1	Estadificación clínica: cT-3N0M0 (EII)	Estadificación clínica: y el Chatbot cT3N1M0 (EIII)
3	Tratamiento	Sexo femenino, 63 años, HTA y trasplante renal Carcinoma escamoso de tercio medio de esófago	CROSS y esofagectomía	Quimioterapia Radioterapia definitiva
4	Tratamiento	Sexo femenino, 80 años, ingresa por Guardia por hemorragia digestiva y disnea. Hallazgo de adenocarcinoma gástrico cT3N+M0	Prehabilitación 3 semanas y gastrectomía	Tratamiento paliativo
5	Tratamiento	Masculino, 65 años, cáncer gástrico difuso (linnitis), mala tolerancia a la vía oral. Sin cobertura social	Prehabilitación 3 semanas internado (nutrición parenteral y enteral) y gastrectomía	FLOT y gastrectomía
6	Tratamiento	Sexo femenino, 51 años, adenocarcinoma de la UEG cT3N+M0Discordancia entre Siewert	FLOT y gastrectomía	CROSS y esofagectomía
7*	Tipo de cirugía	Paciente con adenocarcinoma de cuerpo gástrico	FLOT y gastrectomía total	FLOT y gastrectomía distal

*Este caso no se considera discrepancia formal (el plan terapéutico es correcto), pero se menciona ya que varía el tipo de gastrectomía, lo que es relevante revisar para asegurar una resección oncológica adecuada.
UEG: unión gastroesofágica; PET: tomografía por emisión de positrones (PositronEmissionThomography); SUV: StandardizedUptakeValue; HTA: hipertensión arterial; CROSS: ChemoRadiotherapy for Oesophageal cancer followed by Surgery; FLOT: quimioterapia perioperatoria con 5-Fluorouracilo, leucovorina, oxaliplatino y tetraxetaxel (docetaxel).

podría contribuir a estrategias terapéuticas personalizadas.

Un estudio reciente de China[15] comparó el rendimiento de ChatGPT-4o y Gemini Advanced® en la toma de decisiones para cáncer gástrico avanzado, utilizando tres enfoques complementarios: preguntas clínicas clave, pacientes evaluados en reuniones de equipo multidisciplinario y casos raros publicados en PubMed. En todos los escenarios, ChatGPT-4o mostró un desempeño superior, con mayor precisión y completitud de las recomendaciones, mejor alineamiento con guías clínicas y mayor capacidad para adaptarse a perfiles complejos de pacientes. La tasa de alucinaciones también fue menor en ChatGPT-4o (5,25%) en comparación con Gemini (8,42%), sin implicar riesgos clínicos inmediatos. Si bien estos resultados refuerzan el potencial de los modelos conversacionales como apoyo a la toma de decisiones, el propio estudio destaca que, debido a las limitaciones inherentes de las IA y la ausencia de guías formales de uso, estas herramientas deben integrarse únicamente bajo supervisión experta y en el marco de contextos clínicos reales.

Es importante señalar que ChatGPT no es un sistema de IA diseñado específicamente para uso médico, ni cuenta con certificación como dispositivo médico. Se trata de un modelo de lenguaje generalista, cuya aplicación en el ámbito sanitario debe realizarse bajo supervisión profesional y no sustituye el juicio clínico especializado. Sin embargo, su amplia disponibilidad, fácil acceso y utilización, lo convierten en una herramienta atractiva para explorar su potencial como asistente en la práctica clínica real, especialmente en contextos donde no se dispone de sistemas de IA médica específicos.

La evaluación clínica inicial del cáncer gástrico o esofágico permite clasificar a los pacientes en cuatro grupos[7,8].

1) tumores tempranos, resecables por endoscopia;
2) tumores pequeños no endoscópicamente resecables, sin adenopatías ni metástasis, candidatos a cirugía primaria;
3) enfermedad localmente avanzada, con invasión profunda y/o adenopatías locorregionales, que se beneficia de tratamiento neoadyuvante o perioperatorio previo a la cirugía, y
4) enfermedad metastásica o con adenopatías fuera del territorio regional (estadio IV), cuya conducta es paliativa.

Una correcta clasificación, junto con el estado general y las comorbilidades, orienta la selección del tratamiento más adecuado.

En nuestra serie, tanto el equipo médico como el *chatbot* enfrentaron dificultades para discriminar entre T2-T3 y T3-T4a en algunos casos, dado que la interpretación dependió exclusivamente de los informes de TC. Esto se debe a que, si bien la TC es el método más utilizando en la práctica cotidiana para estadificar cáncer gástrico [16], como los tumores T2 y T3/4 se reconocen por signos transmurales, la diferenciación puede ser difícil, particularmente entre T3 y T4a, donde la serosa no se visualiza directamente y el tejido graso perigástrico varía entre pacientes. Algunos hallazgos como irregularidad de la superficie externa, opacidad de la grasa perigástrica o hiperatenuación serosa sugieren T4a, mientras que la extensión franca a órganos adyacentes orienta a T4b[17,18].

La evaluación del compromiso ganglionar es clave para definir la necesidad de terapia neoadyuvante, aunque no existe un método de referencia (gold standard) en cáncer gástrico y de la UEG. La ecoendoscopia presenta una exactitud variable (30-90%), con baja sensibilidad en estaciones 7-12 y marcada dependencia del operador, lo que limita su disponibilidad[19]. La tomografía computarizada tiene una tasa considerable de falsos negativos, especialmente en tumores difusos o mixtos, y utiliza criterios morfológicos (tamaño, forma y realce), siendo el tamaño el más relevante.

La PET-TC aporta poco en tumores gástricos difusos debido a la baja avidez por FDG, dificultando aún más la estadificación ganglionar[20]. Para metástasis a distancia, la TC continúa siendo el estudio de elección, aunque su sensibilidad y especificidad para enfermedad peritoneal son moderadas (66% y 77%). Los hallazgos sugestivos incluyen ascitis, nódulos o placas peritoneales, engrosamiento sobre intestino delgado o hiperrealce del peritoneo.

En nuestra práctica, la estadificación se realiza con TC con contraste oral e intravenoso. No contamos con ecoendoscopia y la PET-TC no se solicita de rutina en tumores gástricos, aunque algunos pacientes la aportan desde otras instituciones; en contraste, sí la utilizamos con mayor frecuencia en cáncer de esófago y de la UEG.

Respecto a la estadificación clínica del cáncer de esófago, la ecoendoscopia parece ser precisa para el T, con una sensibilidad y especificidad del 85 y 87% para predecir el T1a. En cuanto a los ganglios linfáticos, la precisión es más variable, pero puede mejorarse realizando punción con aguja fina de los ganglios linfáticos[21].

La tomografía computarizada es el estudio más utilizado para evaluar el cáncer de esófago, pero tiene una baja sensibilidad en los estadios tempranos, y una baja sensibilidad y especificidad para evaluar los ganglios linfáticos, con falsos negativos en casos de micrometástasis, y de falsos positivos en presencia de infecciones o patologías inflamatorias. La PET-TC es más sensible para identificar los tumores primarios, la diseminación ganglionar y metástasis a distancia en este tipo de cáncer, que la tomografía. Sin embargo, la baja resolución espacial genera que esta no sea fiable para evaluar la extensión y profundidad de la lesión (T). Respecto del N (ganglios linfáticos), puede tener pobre identificación en los ganglios adyacentes al tumor, sin

diferenciarlos de él. Además, las micrometástasis pueden no captar el FDG. Mientras que, para el M (metástasis), la PET-TC tiene la mejor performance, con una sensibilidad del 71% y una especificidad del 93%.

En nuestra serie, la estadificación se basó principalmente en TC y PET-TC, dado que la ecoendoscopia no estaba sistemáticamente disponible al momento del diagnóstico. Aun así, el *chatbot* logró una concordancia adecuada con el comité de tumores, lo que sugiere que estas herramientas pueden integrarse a la práctica real incluso cuando no se cuenta con todos los métodos diagnósticos ideales.

Este estudio presenta varias limitaciones. Su diseño retrospectivo implica heterogeneidad y posible pérdida de datos, sumado a la variabilidad en los informes de estudios realizados en distintos centros, lo que dificulta la estandarización. Además, el análisis se basó en informes escritos y no en imágenes originales, lo que impide una revaluación objetiva por parte del modelo conversacional.

Ningún paciente contó con ecoendoscopia (EUS), método de referencia para evaluar profundidad tumoral y compromiso ganglionar, lo que podría haber limitado la precisión del cTNM en algunos casos. Sin embargo, la baja disponibilidad de esta tecnología en la Argentina, hace que los resultados sean pragmáticos y extrapolables a la mayoría de los centros locales. Asimismo, al tratarse de un estudio de un único centro de tercer nivel, los resultados no son generalizables sin validación externa, requisito indispensable para modelos de IA[22]. Finalmente, el tamaño muestral reducido limita el poder estadístico y la detección de patrones infrecuentes.

En conclusión, sobre la base de los resultados presentados, es posible afirmar que la IA, aplicada mediante un *chatbot* entrenado con guías clínicas, demostró utilidad para apoyar la estadificación y la planificación terapéutica inicial del cáncer esofagogástrico. Aun con las limitaciones del estudio, se observó una concordancia elevada con las decisiones del equipo tratante. Estos resultados posicionan a los modelos conversacionales como herramientas prometedoras y accesibles, cuya integración debe continuar evaluándose en estudios prospectivos y multicéntricos.

■ ENGLISH VERSION

Introduction

Artificial Intelligence (AI) has emerged as a transformative medical tool, particularly in oncology, demonstrating its ability to improve diagnostic accuracy, optimize treatment planning, and support personalized decision-making[1]. Broadly speaking, AI refers to systems capable of replicating human cognitive functions, such as learning and reasoning[2]. In this field, natural language models enable conversational assistants—chatbots—to interpret and interact with complex clinical language.

Advances in deep learning have expanded the ability of these technologies to analyze diverse information—including images, free-text data, and structured clinical data—and link it to clinical practice guidelines. In the field of gastrointestinal oncology, notable successes have been made in the early detection of gastroesophageal cancer, the automated identification of endoscopic lesions, and the prediction of treatment response[3]. However, little evidence exists regarding the use of conversational models to develop assistants that can help with staging and initial treatment planning.

In Argentina, evidence regarding the use of AI in medicine is still limited. Studies have been published on the use of AI to identify critical view of safety during cholecystectomy procedures at Hospital Argerich[4], of ChatGPTforcolorectalcancerscreeningrecommendations[5], and on implementing an AI co-pilot integrated into electronic medical records at Hospital Italiano de Buenos Aires[6] to improve access to relevant patient information.

In this context, the primary objective of our study was to assess the accuracy of an AI-driven conversational model in predicting clinical staging (cTNM and overall stage) using diagnostic data available at the initial oncology consultation for patients with gastroesophageal cancer. The secondary objective was to analyze the concordance between the treatment recommendations suggested by the chatbot and those provided by the tumor board or the treating team.

Material and methods

We conducted a retrospective study using the database of patients who had their first consultation for a cancer-related condition at the institution's gastroesophageal surgery clinic between December 2023 and May 2025.

The study included patients with complete preoperative workups, defined as patients diagnosed with esophageal, gastroesophageal junction (GEJ), or gastric tumors, an upper gastrointestinal endoscopy report, a biopsy with pathology report, computed tomography (CT) scan, and positron emission tomography-computed tomography (PET-CT) in patients with esophageal cancer or tumors of the GEJ.

Patients who did not undergo all the preoperative tests mentioned or whose histology was other than adenocarcinoma or squamous cell carcinoma were excluded.

Treatment decisions were made by a multidisciplinary team of esophageal cancer surgeons,

clinical oncologists, anesthesiologists, general practitioners, and nutritionists.

Development and use of the virtual assistant

For automated analysis, a virtual clinical assistant was developed using ChatGPT (OpenAI), version GPT-4.5®, via a subscription to ChatGPT Plus®. The development and configuration of the assistant were carried out by surgeons trained in gastroesophageal surgery and with expertise in tumor staging. Although these surgeons lack formal training in programming or algorithm development, they have received training in AI related to the healthcare field. Therefore, they were able to structure the use of the model in accordance with appropriate clinical and methodological criteria.

The chatbot was specifically configured for this task through data ingestion and contextual parsing of the updated clinical guidelines from the National Comprehensive Cancer Network (NCCN)[7,8] and the European Society for Medical Oncology (ESMO)[9,10] on esophageal and gastric cancer.

The model was trained to process each patient's clinical information in natural language, simulating an initial oncology consultation. It should be noted that, for this purpose, the algorithm was not retrained or modified. We used prompt engineering to define structured instructions based on the aforementioned international guidelines. Based on the provided data (age, sex, comorbidities, reason for consultation, endoscopy reports, pathology reports, and imaging results), the assistant was required to: a) determine the clinical stage of the tumor according to the 8th edition of the AJCC staging manual[11]; and b) suggest the appropriate initial cancer treatment in accordance with international guidelines.

Prior to the final analysis, pilot tests were conducted using historical cases excluded from the final cohort to standardize data entry formats, refine instructional prompting, and ensure response consistency. There were no substantial changes to the instructions during the analysis of the cases included in the study.

Each case study was processed individually, uploading the clinical data to the platform and recording the AI-generated responses in a structured table. The recommendations made by the assistant were then compared with the decisions made by the tumor board or the treating team in routine clinical practice.

In cases where discrepancies in staging or treatment were identified, targeted interactions with the chatbot were performed to identify their cause. We assessed whether these arose from a valid alternative interpretation, an omission of relevant data, or a limitation in the model's contextual understanding of the test results.

Statistical analysis

The variables analyzed included:

1) demographic data: age, sex, comorbidities and reasons for consultation; 2) tumor data: location, histopathologic characteristics, endoscopy reports, CT-scan and PET-CT results, cTNM, clinical stage; 3) treatment: endoscopic treatment, surgery without neoadjuvant therapy, surgery with neoadjuvant therapy or perioperative chemotherapy, palliative treatment.

A descriptive analysis was conducted using the data collected. Categorical variables were expressed as absolute frequencies and percentages and continuous variables as mean and range.

Concordance between the clinical staging (cTNM and overall stage) proposed by the chatbot and that determined by the medical team was assessed via direct case-by-case comparison. Similarly, the concordance in the suggested therapeutic approach was analyzed.

The study was designed based on the hypothesis that an AI conversational model, trained using updated clinical guidelines, could correctly predict clinical staging in at least 80% of cases and suggest treatments in line with international recommendations. Given the exploratory nature of the study and the sample size, inferential statistical tests were not performed.

The accuracy of the model was calculated by dividing the number of cases where the chatbot's response matched the medical team staging or therapeutic strategy by the total number of evaluated cases. The following formula was used:

Accuracy (%) = (Concordant cases / Total cases) x 100

Accuracy was calculated separately for cTNM, clinical stage, and recommended treatment.

Ethical considerations: The study was conducted following the recommendations of the Declaration of Helsinki. As the study was observational and based on the analysis of retrospective anonymous clinical data, an informed consent was not required

Results

Of the 53 patients who sought medical care for esophageal, gastric, or GEJ tumors during the specified period, 19 were excluded due to an absence of relevant data in their medical records, resulting in 34 cases that met the inclusion criteria. Mean age was 67 years (range 51-80) and 51% (n = 18) were men. The reason for consultation was documented in 29 patients and is presented in Table 1, along with the other demographic data and tumor location.

Three patients were in stage I, 7 in stage II, 17 in

TABLE 1

Demographic data and tumor characteristics

Variables	n (%)
Comorbidities:	
Diabetes mellitus	11 (32%)
HTN	19 (56%)
COPD	6 (18%)
Current or former smoker	9 (27%)
Cardiovascular disease	9 (27%)
Breast cancer	6 (18%)
Other cancers	4 (11%)
Reason for consultation	
Gastrointestinal bleeding	6 (18%)
Dysphagia	16 (47%)
Dyspepsia or epigastric pain	14 (41%)
Pyloric syndrome or vomiting	9 (26%)
Weight loss	
Tumor location	
Esophagus	10 (29%)
Stomach	13 (38%)
Gastroesophageal junction	11 (32%)
- Siewert I	2 (18%)
- Siewert II	6 (54%)
- Siewert III	3 (27%)
Anatomical pathology	
Signet ring cells	
Esophagus	1 (10%)
Gastroesophageal junction	3 (27%)
Stomach	4 (31%)
Squamous cell carcinoma*	5 (50%)

*Of total esophageal cancers.
COPD: chronic obstructive pulmonary disease: HTN: hypertension.

TABLE 2

Comparison between clinical staging determined by the medical team and the chatbot

cTNM**	Medical team	Chatbot
cT		
cT1b	1(3%)	1 (3%)
cT1b-2	1(3%)	1 (3%)
cT2-3	3(9%)	3 (9%)
cT3	21(62%)	21 (62%)
cT3-4	4(12%)	4 (12%)
cT4	4(12%)	4 (12%)
cN		
cN0-1	1(3%)	1 (3%)
cN+	20(59%)	21 (62%)
cM		
cM1	7(21%)	6(18 %)
Stage		
I	3(9%)	3 (9%)
II	7(20%)	6 (18%)
III	17(50%)	19 (56%)
IV	7(20%)	6 (18%)

stage III, and 7 in stage IV. The rest of the characteristics are described in Table 2.

The accuracy between the cTNM staging determined by the medical team and that of the chatbot was 94%. Concordance of the treatment proposed was 85%.

Discordant cases are presented in Table 3.

When the cohort was divided into three sequential groups (n = 12, 11, and 11 patients, respectively), the accuracy of the chatbot to predict cTNM and staging improved to 83% in group 1, 100% in group 2, and 100% in group 3.

Discussion

The present study evaluated the usefulness of an AI conversational model, trained using up-to-date clinical guidelines, to assist with staging and initial treatment planning in patients with esophageal, gastroesophageal junction, and gastric cancer. The results show a high level of concordance (94%) between the clinical staging proposed by the chatbot and that determined by the treating medical team, both in terms of cTNM and overall stage. Although concordance in therapeutic management was somewhat lower (85%), the differences were mostly explained by contextual clinical considerations that were not always available to the model, such as social constraints.

Chatbots are programs that process and simulate human conversations using natural language[12]. The use of chatbots in healthcare increased during the COVID-19 pandemic, although it remains underutilized and is expected to increase over the next decade. In the field of gastric cancer, various artificial intelligence models have demonstrated promising accuracy in identifying high-risk individuals and predicting tumor aggressiveness, recurrence, and survival[13,14]. Deep learning has demonstrated the ability to analyze endoscopic images, genomic data, and histopathological data, thereby improving diagnosis and prognosis. Integrating multimodal clinical, genetic, and imaging data provides a comprehensive disease profile, facilitating the implementation of personalized treatment strategies.

A recent study from China[15] compared the performance of ChatGPT-4o and Gemini Advanced® in decision-making for advanced gastric cancer. The study utilized three complementary approaches: key clinical questions, patients evaluated in multidisciplinary team meetings, and rare cases published in PubMed. ChatGPT-4o demonstrated superior performance in all scenarios, providing more accurate and comprehensive recommendations, better aligning with clinical guidelines, and demonstrating a greater ability to adapt to complex patient profiles. The hallucination rate was also lower in ChatGPT-4o (5.25%) compared to Gemini (8.42%), without implying immediate clinical risks. While

■ TABLE 3

Discordant cases between the medical team and the chatbot

Case	Discrepancy category	Reason	Medical team	chatbot
1	cTNM and clinical stage	Lateral aortic lymph nodes on imaging studies	Considers M1, stage IV	Considers regional lymph nodes, M0, stage III
1	Treatment	Discrepancy of clinical stage (IV vs. III)	Palliative treatment	Neoadjuvant therapy and surgery
2	cTNM and clinical stage	GEJ tumor, without uptake in lymph nodes on PET-CT. The chatbot considers it to be N+ based on the tumor's length (5.8 cm) and SUV of 13.1	Clinical stage: cT3N0M0 (EII)	Clinical stage: and chatbot cT3N1M0 (EIII)
3	Treatment	Female, 63 years old, HTN, and kidney transplant; squamous cell carcinoma of the middle third of the esophagus	CROSS and esophagectomy	Chemotherapy Definitive radiation therapy
4	Treatment	Female, 80 years old, admitted to the emergency department with gastrointestinal bleeding and dyspnea. Gastric adenocarcinoma cT3N+M0	Prehabilitation for 3 weeks and gastrectomy	Palliative treatment
5	Treatment	Male, 65 years old, diffuse gastric cancer (linitis plastica). Intolerance to oral feeding. Absence of social security coverage.	Admission for prehabilitation for 3 weeks (parenteral and enteral nutrition) and gastrectomy	FLOT and gastrectomy
6	Treatment	Female, 51 years old, adenocarcinoma of the GEJ cT3N+M0 Discrepancy wirh Siewert	FLOT and gastrectomy	CROSS and esophagectomy
7*	Type of surgery	Gastric body adenocarcinoma	FLOT and total gastrectomy	FLOT and distal gastrectomy

*This case is not considered a formal discrepancy (the treatment plan is correct), but it is mentioned because the type of gastrectomy differs, which is important to review to ensure adequate oncological resection.
GEJ: gastroesophageal junction; PET: positron emission tomography; SUV: standardized uptake value; HTM: hypertension; CROSS: chemoradiotherapy for esophageal cancer followed by surgery; FLOT: perioperative chemotherapy with 5-fluorouracil, leucovorin, oxaliplatin, and docetaxel.

these findings underscore the potential of conversational models to support decision-making, the study also highlights that, due to the inherent limitations of AI and the lack of formal guidelines for their use, these tools should be implemented only under expert supervision and within the context of real-world clinical settings.

It is noteworthy that ChatGPT is not an AI system designed specifically for medical use, nor is it certified as a medical device. It is a general-purpose language model; its use in the healthcare setting must be carried out under professional supervision and does not replace specialized clinical judgment. However, its high availability, accessibility, and ease of use make it an attractive tool for exploring its potential in real-world clinical practice, particularly where specialized medical AI systems are unavailable.

Initial clinical evaluation of gastric or esophageal cancer allows for patient classification into four groups[7,8]:

1) early-stage tumors, amenable to endoscopic resection;
2) localized tumors not suitable for endoscopic resection, without lymph node involvement or metastasis, eligible for primary surgery;
3) locally advanced disease, characterized by deep invasion and/or locoregional lymph node involvement, benefiting from neoadjuvant or perioperative treatment prior to surgery; and
4) metastatic disease or lymph node involvement extending beyond the regional territory (stage IV), where management is primarily palliative.

An accurate staging, along with the patient's performance status and comorbidities, guides the selection of the most appropriate treatment.

In our series, both the medical team and the chatbot encountered difficulties in distinguishing between T2-T3 and T3-T4a in some cases, since the interpretation relied solely on CT reports. Although CT is the most commonly used method for staging gastric cancer[16], T2 and T3/4 tumors are identified by transmural features. Consequently, differentiation can be challenging, especially between T3 and T4a tumors, as the serosa is not directly visualized and the amount of perigastric fat varies among patients. Certain findings, such as an irregular outer layer of the gastric wall, haziness of the perigastric fat and the hyperattenuating serosa sign suggest T4a, whereas clear invasion of adjacent organs suggests T4b[17,18].

Assessment of lymph node involvement is key to determining the need for neoadjuvant therapy, although there is no gold standard for gastric and EGJ cancer. The accuracy of endoscopic ultrasound is variable (30–90%), with low sensitivity for stations 7–12 and significant operator dependence, which limits its availability[19]. Computed tomography has a significant false-negative rate, particularly in diffuse or mixed-type tumors. It relies on morphological criteria (size, shape, and enhancement), with size remaining the primary diagnostic factor.

PET-CT provides little information in diffuse gastric tumors due to their low FDG uptake, thereby further complicating lymph node staging[20]. For

detecting distant metastases, CT remains the standard imaging modality, although its sensitivity and specificity for peritoneal involvement are moderate (66% and 77%, respectively). Findings suggestive of peritoneal seeding include ascites, peritoneal nodules or plaques, bowel wall thickening, or peritoneal hyperenhancement.

In our practice, we used oral and intravenous contrast-enhanced CT scans for staging. Endoscopic ultrasound is not available in our setting, and we do not routinely request PET-CT scans for gastric tumors, although some patients bring results from other institutions. We do request PET-CT scans more frequently for esophageal and GEJ cancers.

Endoscopic ultrasound demonstrates high accuracy for staging esophageal cancer T-status, with a sensitivity and specificity of 85% and 87%, respectively, for identifying T1a lesions. The accuracy of endoscopic ultrasound for detecting lymph node involvement is more variable; however, it can be significantly enhanced by incorporating fine-needle aspiration[21].

While CT is the most widely used imaging modality for esophageal cancer, it lacks sensitivity for early-stage disease and demonstrates poor diagnostic performance for lymph node assessment, with high false-negative rates in the presence of micrometastases and false-positive rates due to concurrent infection or inflammation. PET-CT is more sensitive for identifying primary tumors, lymph node involvement, and distant metastases in this type of cancer. However, the low spatial resolution makes it unreliable for assessing the extent and depth of the lesion (T). Regarding N-staging (lymph nodes), PET-CT may face challenges in detecting lymph nodes in close proximity to the primary tumor, as it often fails to distinguish them from the primary lesion. In addition, micrometastases may not uptake FDG. PET-CT scans perform best for M-staging (metastasis), with a sensitivity of 71% and a specificity of 93%.

In our series, staging was primarily based on CT and PET-CT scans, since endoscopic ultrasound was not routinely available at the time of diagnosis. Even so, the chatbot achieved adequate concordance with the tumor board, suggesting that these tools can be incorporated into real-world practice even when not all the ideal diagnostic methods are available.

This study has some limitations. Its retrospective design suggests heterogeneity and the possibility of data loss, along with variability in the reporting of studies conducted at different centers, which hinders standardization. Furthermore, the analysis was based on written reports rather than original images, which prevents the conversational model from conducting an objective reassessment.

No patients underwent endoscopic ultrasound (EUS), the gold standard for assessing tumor depth and lymph node involvement, which may have limited the accuracy of the cTNM staging in some cases. However, the limited availability of this technology in Argentina means that the results are pragmatic and can be extrapolated to most local centers. Furthermore, since this is a single-center tertiary-level study, the results cannot be generalized without external validation, which is an essential requirement for AI models[22]. Finally, the small sample size limits the statistical power and the ability to detect rare patterns.

In conclusion, based on the results presented, it can be stated that AI, implemented through a chatbot trained using clinical guidelines, proved to be useful in supporting staging and initial treatment planning for gastroesophageal cancer. Despite the mentioned limitations, the study demonstrated high concordance with the treatment decisions made by the treating medical team. These findings position conversational models as promising and accessible tools, and their integration should be further evaluated in prospective, multicenter studies.

Referencias bibliográficas /References

1. Marra A, Morganti S, Pareja F, Campanella G, Bibeau F, FuchsT, et al. Artificial intelligence entering the pathology arena in oncology: current applications and future perspectives. Ann Oncol. 2025;36(7):712-25. doi:10.1016/j.annonc.2025.03.006.
2. Chong PL, Vaigeshwari V, Mohammed Reyasudin BK, Noor Hidayah BRA, Tatchanaamoorti P, Yeow JA, et al. Integrating artificial intelligence in healthcare: applications, challenges, and future directions. Future Sci OA. 2025;11(1):2527505. doi:10.1080/20565623.2025.2527505.
3. Chen ZL, Wang C, Wang F. Revolutionizing gastroenterology and hepatology with artificial intelligence: From precision diagnosis to equitable healthcare through interdisciplinary practice. World J Gastroenterol. 2025;31(24):108021. doi:10.3748/wjg.v31.i24.108021.
4. Petracchi EJ, Olivieri SE, Varela J, Canullán CM, Zandalazini H, Ocampo C, et al. Use of artificial intelligence in the detection of the critical view of safety during laparoscopic cholecystectomy. J Gastrointest Surg. 2024;28(6):877-9. doi:10.1016/j.gassur.2024.03.018.
5. Pereyra L, Schlottmann F, Steinberg L, Lasa J. Colorectal Cancer Prevention: Is Chat Generative Pretrained Transformer (Chat GPT) ready to Assist Physicians in Determining Appropriate Screening and Surveillance Recommendations?.J Clin Gastroenterol. 2024;58(10):1022-7. doi:10.1097/MCG.0000000000001979.