Rev Argent Cir. 2026;118(2):e-ededc02- http://dx.doi.org/10.25132/raac.v118.n2.ededc02

Exactitud, estadificación y decisión: la inteligencia artificial frente al cáncer esofagogástrico

Accuracy, staging, and decision-making: artificial intelligence in gastroesophageal cancer

*Enrique Díaz Cantón**

En el cáncer esofagogástrico, la estadificación no es un ejercicio descriptivo: gobierna decisiones de alto impacto y, con frecuencia, divergentes. El trabajo *Exactitud de la inteligencia artificial en la estadificación y toma de decisiones en cáncer esófago-gástrico*[1] formula la pregunta pertinente –cuán exacta es la inteligencia artificial (IA)– y, sobre todo, obliga a precisar contra qué patrón se mide esa exactitud.

La evidencia más directa es alentadora pero condicional. Yao y cols. compararon tres grandes modelos de lenguaje (LLM) de despliegue local con clínicos en la estadificación TNM preoperatoria del cáncer de esófago a partir de informes radiológicos en texto libre, tomando la patología posquirúrgica como patrón de referencia. Al ensayar tres estrategias de consigna –*zero-shot, chain-of-thought* y un método de razonamiento interpretable–, observaron que la consigna estructurada no solo elevó la exactitud, sino que aportó una traza de razonamiento auditable[2]. La IA puede, entonces, alcanzar una exactitud apreciable; pero esa exactitud depende de cómo se la interroga y debe contrastarse contra el oro patológico, no contra la verosimilitud de un texto bien redactado.

Por qué el cáncer esofagogástrico exige cautela adicional

Pocos escenarios castigan tanto un error de estadificación. La conducta depende de distinciones sutiles –el epicentro tumoral respecto de la unión esofagogástrica, la clasificación de Siewert, la divergencia histológica entre adenocarcinoma y carcinoma escamoso– que orientan esquemas perioperatorios o de quimiorradioterapia diferentes. A ello se suman biomarcadores que condicionan la terapia sistémica, como HER2, PD-L1, la inestabilidad de microsatélites y CLDN18.2[3]. Un modelo que confunda un tumor de cardias con uno de la unión, o que omita un biomarcador, propaga el error a todo el plan terapéutico. La "exactitud" que reclama este trabajo debe medirse, por lo tanto, dentro del marco de las guías vigentes y contra el diagnóstico patológico.

De la estadificación a la decisión

Cuando se pasa de estadificar a recomendar, la concordancia se vuelve más variable. En la predicción de conductas de junta tumoral se han descripto congruencias de línea de tratamiento de hasta el 86% con algunos modelos[4], pero el mismo ecosistema produce un tercio de recomendaciones no concordantes cuando se lo interroga sobre tratamiento oncológico abierto, con un 12,5% de alucinaciones[5]. La tabla 1 resume esta evidencia heterogénea, que conviene leer con prudencia.

La conclusión no es desestimar la herramienta, sino encuadrarla. La inteligencia artificial puede afinar la estadificación y aportar un razonamiento explícito y revisable, sobre todo cuando se la guía con consignas estructuradas. Pero la responsabilidad de la decisión –ética, legal y humana– permanece indelegablemente en el cirujano y el oncólogo, que deben validar cada salida contra el paciente concreto, su patología y las

■ TABLA 1

Exactitud y concordancia de la IA en estadificación y decisión oncológica

Dominio evaluado	Modelo(s)	Resultado principal	Fuente
Estadificación TNM de cáncer de esófago desde informes radiológicos (referencia: patología)	INF-72B, Qwen2.5-72B, LLaMA3.1-70B	Exactitud elevada; el razonamiento estructurado (CoT / razonamiento interpretable) superó al zero-shot	[2]
Predicción de recomendaciones de junta tumoral (cabeza y cuello)	Varios LLM	Congruencia de línea de tratamiento hasta 86 %; recomendaciones justificables hasta 98 %	[4]
Recomendaciones oncológicas abiertas vs. guías NCCN (mama, próstata, pulmón)	GPT-3.5®	34,3 % con ≥ 1 recomendación no concordante; 12,5 % "alucinaciones"	[5]

Nota: los valores provienen de estudios con metodologías, idiomas y poblaciones distintas y no constituyen una comparación directa entre modelos. CoT: chain-of-thought; NCCN: National Comprehensive Cancer Network; TNM: tumor-ganglio-metástasis.

**Oncólogo clínico, Instituto Universitario CEMIC, Buenos Aires, Argentina.*
Codirector, Programa de Posgrado en Inteligencia Artificial y Medicina, Academia Nacional de Medicina, Buenos Aires, Argentina.
Correspondencia: ediazcanton@iuc.edu.ar

guías. Medida así, con rigor y contra el patrón correcto, la exactitud de la IA deja de ser una promesa para convertirse en una herramienta; medida contra su propia elocuencia, sigue siendo un espejismo.

Declaración sobre el uso de inteligencia artificial: *El autor declara que en la preparación de este editorial utilizó herramientas de inteligencia artificial generativa basadas en modelos de lenguaje de gran escala como apoyo en la búsqueda y el cotejo de la bibliografía, la organización del texto y la edición de estilo. Todas las afirmaciones, las referencias y los datos cuantitativos fueron verificados por el autor contra las fuentes primarias citadas. El contenido conceptual, las interpretaciones y las conclusiones son de su exclusiva responsabilidad; la inteligencia artificial no figura como autora ni asume responsabilidad alguna sobre el contenido.*

■ ENGLISH VERSION

In gastroesophageal cancer, staging is not merely a descriptive exercise: it guides high-impact, often conflicting decisions. The paper "Accuracy of artificial intelligence for staging and decision-making in esophageal and gastric cancer"[1] raises the pertinent question of how accurate artificial intelligence (AI) is. More importantly, it requires specifying the standard against which this accuracy is measured.

The most direct evidence is encouraging, though conditional. Yao et al. compared the performance of three locally deployed LLMs with that of clinicians in preoperative esophageal cancer staging using free-text radiology reports. The gold standard was derived from postoperative pathological staging. When the authors tested three prompting strategies—zero-shot, chain-of-thought, and an interpretable reasoning method—they observed that the structured prompt not only improved accuracy but also provided an auditable reasoning trace[2]. Therefore, the accuracy of AI can be contingent on the method of querying, and it should be evaluated against the pathological gold standard rather than the plausibility of a well-written text.

Why does gastroesophageal cancer demand particular caution?

Few scenarios impose such a heavy penalty for a staging error. The management of the disease depends on subtle distinctions—tumor location relative to the gastroesophageal junction, Siewert classification, and histological differences between adenocarcinoma and squamous cell carcinoma—that guide different perioperative or chemoradiotherapy regimens. In addition, certain biomarkers, such as HER2, PD-L1, microsatellite instability, and CLDN18.2, condition systemic therapy[3]. A model confounding a cardia tumor with that of the gastroesophageal junction—or one that ignores a biomarker—spreads the error throughout the entire treatment plan. The "accuracy" claimed by this study must therefore be assessed within the framework of current guidelines and in relation to the pathological diagnosis.

From staging to decision-making

When the focus shifts from staging to recommending, concordance becomes more variable. For predicting tumor board procedural recommendations, treatment line congruence of up to 86% has been reported with some models[4], but the same ecosystem generates one-third of non-concordant recommendations when queried about open oncology treatment, with 12.5% of responses being hallucinations[5]. Table 1 summarizes this heterogeneous evidence, which should be read with caution.

The objective is not to discredit the tool, but rather to contextualize its use. Artificial intelligence can improve staging and provide explicit, verifiable reasoning, especially when guided by structured prompts. However, the ethical, legal, and human responsibility for the decision cannot be delegated; it remains with the surgeon and the oncologist, who must verify each output against the individual patient, their pathology, and current guidelines. When evaluated with such rigor

■ **TABLE 1**

Accuracy and concordance of AI in cancer staging and decision-making.

Domain	Model(s)	Primary outcome	Source
TNM staging of esophageal cancer based on radiology reports (gold standard: pathology)	INF-72B, Qwen2.5-72B, LLaMA3.1-70B	High accuracy; structured reasoning (CoT / interpretable reasoning) outperformed zero-shot	[2]
Prediction of tumor board procedural recommendations (head and neck)	Various LLMs	Treatment line congruence of up to 86%; justifiable recommendations up to 98%.	[4]
Open oncology recommendations vs. NCCN guidelines (breast cancer, prostate cancer, lung cancer)	GPT-3.5	34.3% with ≥ 1 nonconcordant option; 12.5% "hallucinations"	[5]

Note: These values are derived from studies using different methodologies, languages, and populations and do not constitute a direct comparison between models. CoT: chain-of-thought; NCCN: National Comprehensive Cancer Network; TNM: tumor-node-metastasis

and against the proper standard, the accuracy of AI shifts from being a promise to becoming a tool; measured against its own eloquence, it remains a mirage.

Artificial intelligence disclosure: *The author declares that, during the preparation of this editorial, generative artificial intelligence tools based on large language models were utilized to assist with literature search and cross-referencing, text organization, and copyediting. All statements, references, and quantitative data were verified by the author against the primary sources cited. The conceptual content, interpretations, and conclusions are the sole responsibility of the author; the artificial intelligence is not credited as an author, nor does it assume any responsibility for the content.*

Referencias bibliográficas /References

1. Coturel AE, Pereyra P, García R, Diomedi A, Cabas Audicio JJM, Klappenbach R. Exactitud de la inteligencia artificial en la estadificación y toma de decisiones en cáncer esófago gástrico. Rev Argent Cir. 2026;118(2):e-1963. DOI: http://dx.doi.org/10.25132/raac.v118.n2.1963
2. Yao Y, Cen X, Gan L, Jiang J, Wang M, Xu Y, Yuan J. Automated Esophageal Cancer Staging From Free-Text Radiology Reports: Large Language Model Evaluation Study. JMIR Med Inform. 2025;13:e75556. doi:10.2196/75556.
3. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines): Gastric Cancer y Esophageal and Esophagogastric Junction Cancers. NCCN; 2025-2026. Disponible en NCCN.org.
4. Aubreville M, Ganz J, Ammeling J, et al. Prediction of tumor board procedural recommendations using large language models. Eur Arch Otorhinolaryngol. 2025;282(3):1619-29. doi:10.1007/s00405-024-08947-9.
5. Chen S, Kann BH, Foote MB, Aerts HJWL, Savova GK, Mak RH, Bitterman DS. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. JAMA Oncol. 2023;9(10):1459-62. doi:10.1001/jamaoncol.2023.2954.