

VALIDACIÓN DE LA CONCORDANCIA EXPERTO-ALGORITMO PARA LA ESTRATIFICACIÓN INICIAL DEL RIESGO Y LA EVALUACIÓN DE LA RESPUESTA DINÁMICA EN CARCINOMA DIFERENCIADO DE TIROIDES

Ramírez Stieben, Luis Agustín;¹ Melado, Graciela Eva;² Smulever, Anabella;³ Abelleira, Erika;⁴ Califano, Inés;⁵ Mollerach, Ana María;⁶ Galíndez, María Macarena;⁷ Damilano, Roxana Analía;⁸ Gonza, María Natalia;⁷ Pitoia, Fabián⁴

1. Servicio de Endocrinología del Grupo Gamma, Rosario (Argentina). 2. Megan SRL, Santa Rosa, La Pampa. Argentina. 3. Instituto de Investigaciones Médicas Alfredo Lanari, Universidad de Buenos Aires. Ciudad de Buenos Aires. Argentina. 4. División Endocrinología, Hospital de Clínicas, Universidad de Buenos Aires. Ciudad de Buenos Aires. Argentina. 5. Servicio de Endocrinología del Instituto de Oncología AH Roffo, Universidad de Buenos Aires. Ciudad de Buenos Aires. Argentina. 6. Hospital Italiano de Buenos Aires. Ciudad de Buenos Aires (Argentina). 7. Servicio de Endocrinología del Hospital Dr. Arturo Oñativía. Salta. Argentina. 8. Sanatorio Allende, Córdoba. Argentina.

INFO ARTÍCULO

Palabras clave:

carcinoma diferenciado de tiroides, estratificación del riesgo, respuesta al tratamiento, algoritmo clínico, herramientas digitales.

RESUMEN

La estratificación del riesgo de recurrencia y la evaluación dinámica de la respuesta son componentes centrales del manejo del carcinoma diferenciado de tiroides (CDT). Su aplicación requiere integrar múltiples variables clínicas, bioquímicas, anatomopatológicas e imagenológicas, lo que puede generar variabilidad entre observadores. En este contexto, las herramientas digitales basadas en algoritmos clínicos preespecificados podrían facilitar una clasificación estandarizada y reproducible. El objetivo de este estudio fue evaluar la concordancia entre un algoritmo clínico digital y la clasificación independiente realizada por expertos en CDT. Materiales y métodos: estudio metodológico de validación mediante 50 casos clínicos sintéticos evaluados de forma independiente por 9 expertos, todos participantes del Consenso Delphi Intersocietario argentino sobre CDT. Se analizaron 450 evaluaciones experto-caso. Los dominios principales fueron el riesgo inicial de recurrencia y la respuesta dinámica al tratamiento; el estadio pronóstico constituyó un dominio secundario. La concordancia se evaluó mediante acuerdo observado, kappa no ponderado, kappa ponderado lineal y coeficiente AC1 de Gwet. Los intervalos de confianza del 95% se estimaron mediante bootstrap por conglomo-

merados a nivel del caso clínico. Resultados: el acuerdo para el riesgo inicial en cuatro categorías fue del 88,4% (kappa 0,840) y aumentó al 91,8% al agruparlo en tres categorías (kappa 0,876). Para la respuesta dinámica, el acuerdo fue del 87,7% (kappa 0,830), y para el estadio pronóstico, del 93,3% (kappa 0,872). Los coeficientes kappas ponderados lineales variaron entre 0,812 y 0,875. Las discrepancias se concentraron predominantemente entre categorías adyacentes; no hubo discordancias directas entre las categorías de bajo y alto riesgo. Conclusiones: el algoritmo mostró elevada concordancia con la clasificación realizada por expertos. Estos resultados respaldan la factibilidad de implementar criterios clínicos complejos mediante una herramienta digital estructurada y justifican su evaluación prospectiva en la práctica clínica real.

Recibido: 20 de julio de 2026
Aceptado: 11 de agosto de 2026

VALIDATION OF EXPERT-ALGORITHM AGREEMENT FOR INITIAL RISK STRATIFICATION AND DYNAMIC RESPONSE ASSESSMENT IN DIFFERENTIATED THYROID CANCER

ARTICLE INFO

Keywords:
differentiated thyroid cancer, risk stratification, response to treatment, clinical algorithm, digital tools.

ABSTRACT

Risk stratification for recurrence and dynamic response assessment are central components of differentiated thyroid cancer (DTC) management. Their application requires the integration of multiple clinical, biochemical, pathological, and imaging variables, which may lead to interobserver variability. In this context, digital tools based on prespecified clinical algorithms could facilitate standardized and reproducible classification. The objective of this study was to evaluate the agreement between a digital clinical algorithm and independent expert classification in DTC. Materials and methods: this methodological validation study used 50 synthetic clinical cases independently assessed by 9 experts, all of whom had participated in the Argentine intersociety Delphi consensus on DTC. Overall, 450 expert-case evaluations were analyzed. The primary domains were initial recurrence risk and dynamic response to treatment; prognostic stage was a secondary domain. Agreement was assessed using observed agreement, unweighted kappa, linearly weighted kappa, and Gwet's AC1 coefficient. Ninety-five percent confidence intervals were estimated through cluster bootstrap resampling at the clinical-case level. Results: agreement for initial risk across four categories was 88.4% (kappa 0.840) and increased to 91.8% when grouped into three categories (kappa 0.876). Agreement was 87.7% for dynamic response (kappa 0.830) and 93.3% for prognostic stage (kappa 0.872). Linearly weighted kappa values ranged from 0.812 to 0.875. Disagreements occurred predominantly between adjacent categories, with no direct discordance between low- and high-risk classifications. Conclusions: The algorithm showed high agreement with expert classification. These findings support the feasibility of implementing complex clinical criteria through a structured digital tool and justify prospective evaluation in real-world clinical practice.

Received: July 20, 2026
Accepted: August 11, 2026

INTRODUCCIÓN

El carcinoma diferenciado de tiroides (CDT) comprende un grupo de neoplasias con un pronóstico generalmente favorable, aunque con una marcada heterogeneidad en el riesgo de persistencia, recurrencia y mortalidad específica. Esta variabilidad ha impulsado la transición desde estrategias terapéuticas relativamente uniformes hacia un enfoque individualizado y adaptado al riesgo, en el cual la estimación pronóstica contribuye a definir la extensión del tratamiento inicial, la indicación de terapias adyuvantes y la intensidad del seguimiento. Las recomendaciones actuales de la *American Thyroid Association* (ATA) consolidan este modelo de manejo basado en una estratificación del riesgo progresiva y dinámica a lo largo de la evolución clínica [1,2].

La evaluación pronóstica del CDT integra sistemas con objetivos diferentes y complementarios. El sistema TNM/AJCC, actualmente en su octava edición, fue diseñado principalmente para estimar el riesgo de mortalidad específica por cáncer, mientras que los sistemas de estratificación inicial permiten estimar el riesgo de persistencia o recurrencia a partir de características clínicas y anatomopatológicas [1,3]. La actualización de las guías ATA de 2025 introdujo una clasificación más granular del riesgo de recurrencia, con el propósito de reflejar mejor la heterogeneidad pronóstica y favorecer decisiones terapéuticas adaptadas al perfil individual de cada paciente [1].

Sin embargo, la estimación pronóstica inicial, sin embargo, constituye solo el punto de partida del seguimiento clínico. La información obtenida durante el seguimiento, incluidos los niveles de tiroglobulina o la evolución de los anticuerpos antitiroglobulina y los hallazgos estructurales en los estudios por imágenes, permite reclasificar el riesgo según la respuesta al tratamiento. El concepto de estratificación dinámica mostró que la evaluación de la respuesta puede modificar de manera sustancial las estimaciones pronósticas iniciales y proporcionar una aproximación más precisa al riesgo futuro de persistencia o recurrencia [4]. Posteriormente, este enfoque fue validado en diferentes contextos terapéuticos,

incluidos pacientes tratados sin radioyodo, y se ha consolidado como un componente central del seguimiento adaptado al riesgo [5,6].

La aplicación de estos sistemas requiere integrar múltiples variables clínicas, anatomopatológicas, bioquímicas y de imágenes, algunas de ellas dependientes de puntos de corte y reglas de combinación específicas. La traducción de estos criterios a algoritmos clínicos preespecificados podría favorecer una aplicación más estandarizada, reproducible y consistente de las recomendaciones clínicas. Sin embargo, antes de su implementación como herramientas de apoyo a la decisión, es necesario demostrar que las clasificaciones generadas reproduzcan de manera confiable las realizadas por médicos expertos. En Argentina, el Consenso Delphi Intersocietario sobre diagnóstico, tratamiento y seguimiento del CDT incorporó y contextualizó recomendaciones para su aplicación en la práctica clínica local [7]. En los últimos años se han desarrollado calculadoras internacionales que trasladan criterios contemporáneos de estratificación a interfaces digitales, como la calculadora ATA 2025 de TIRO [8]. No obstante, persisten diferencias entre las herramientas disponibles en cuanto a los dominios integrados, la lógica de clasificación, la correcta traducción y su adaptación a contextos clínicos regionales [8].

En este sentido, el objetivo del presente estudio fue evaluar la concordancia entre las clasificaciones generadas por un algoritmo clínico digital preespecificado y las realizadas de forma independiente por médicos expertos en CDT para la estratificación inicial del riesgo de recurrencia y la evaluación dinámica de la respuesta al tratamiento. Como objetivo secundario, se evaluó la concordancia para la clasificación pronóstica según el sistema AJCC/TNM (8.ª edición).

MATERIALES Y MÉTODOS

Diseño del estudio

Se realizó un estudio metodológico de validación clínica mediante casos estandarizados, diseñado para evaluar la concordancia entre las clasificaciones generadas por un algoritmo clínico preespecificado y las asignadas de manera

independiente por médicos expertos en CDT. Los dominios principales de validación fueron la estratificación inicial del riesgo de recurrencia y la estratificación dinámica de respuesta al tratamiento. La estadificación AJCC constituyó un dominio secundario de validación.

La herramienta evaluada fue un prototipo de calculadora clínica digital integrado en FASEN Tools, una aplicación web progresiva orientada al apoyo de la toma de decisiones en endocrinología. Para el presente estudio se evaluaron exclusivamente los módulos vinculados con el CDT, diseñados para asistir en la estimación del riesgo inicial de recurrencia, la estadificación TNM/AJCC y la estratificación dinámica de la respuesta al tratamiento. Al momento del estudio, los módulos evaluados se encontraban en fase de prototipo y no estaban disponibles para uso asistencial general.

Los algoritmos clínicos fueron implementados en JavaScript y organizados en módulos independientes para la estratificación inicial del riesgo de recurrencia, la estadificación TNM/AJCC y la estratificación dinámica de respuesta al tratamiento. La herramienta no fue considerada un estándar clínico absoluto, sino un algoritmo preespecificado contra el cual se evaluó la concordancia del juicio experto.

El algoritmo fue construido a partir de criterios derivados de guías internacionales recientes y consensos nacionales, incluyendo las guías ATA 2025, el sistema AJCC/UICC TNM, 8.^a edición, y el Consenso Delphi Intersocietario argentino sobre diagnóstico, tratamiento y seguimiento del CDT.

Desarrollo de los casos clínicos

Se diseñaron 50 casos clínicos sintéticos mediante la construcción deliberada de escenarios estandarizados de CDT. Los casos no correspondieron a pacientes reales, sino que fueron elaborados a partir de combinaciones clínicamente plausibles de variables utilizadas en la estratificación inicial del riesgo, la estadificación AJCC y la evaluación dinámica de la respuesta al tratamiento.

Los escenarios fueron contruidos para representar un amplio espectro de situaciones

clínicas habituales y potencialmente complejas en el CDT, procurando cubrir las principales categorías de riesgo inicial y de respuesta dinámica definidas por el algoritmo. Los casos incluyeron diferentes perfiles histológicos y pronósticos, entre ellos carcinoma papilar clásico y variantes histológicas de mayor agresividad, carcinoma folicular, carcinoma oncocítico, enfermedad ganglionar central y lateral, metástasis a distancia, invasión extratiroidea, invasión vascular, compromiso de márgenes quirúrgicos, diferentes valores de tiroglobulina posoperatoria, presencia de anticuerpos antitiroglobulina y hallazgos estructurales en los estudios por imágenes.

Los escenarios abarcaron las cuatro categorías de riesgo inicial de recurrencia utilizadas por el algoritmo –bajo, intermedio-bajo, intermedio-alto y alto–, así como las cuatro categorías de respuesta dinámica al tratamiento: excelente, indeterminada, bioquímicamente incompleta y estructuralmente incompleta.

Para cada caso se generó, mediante la aplicación del algoritmo preespecificado, una clasificación de referencia que incluyó la categoría de riesgo inicial, el estadio AJCC y la categoría de respuesta dinámica. Esta clasificación fue establecida antes del análisis de las respuestas de los expertos.

Participantes

Se realizó un muestreo intencional de médicos especialistas en endocrinología con experiencia en el manejo del CDT. Todos los participantes ejercían en centros con atención habitual de pacientes con cáncer de tiroides y habían integrado el panel de expertos del Consenso Delphi Intersocietario argentino sobre diagnóstico, tratamiento y seguimiento del CDT. La selección procuró incluir profesionales de diferentes ámbitos asistenciales.

Cada experto recibió los 50 casos clínicos en una planilla estructurada y fue solicitado a clasificar, de acuerdo con su criterio clínico, el riesgo inicial de recurrencia, la estadificación TNM/AJCC y la respuesta dinámica al tratamiento.

Las evaluaciones se realizaron de manera independiente y sin acceso a las clasificaciones

generadas por el algoritmo. Los expertos tampoco tuvieron acceso a las respuestas de los demás participantes. La participación fue voluntaria.

Variables evaluadas

Los dominios principales de validación fueron la estratificación inicial del riesgo de recurrencia y la estratificación dinámica de respuesta al tratamiento.

La estratificación inicial del riesgo de recurrencia se evaluó en cuatro categorías: bajo, intermedio-bajo, intermedio-alto y alto. Asimismo, se realizó un análisis complementario agrupando las categorías intermedio-bajo e intermedio-alto en una única categoría de riesgo intermedio, obteniéndose una clasificación en tres categorías: bajo, intermedio y alto.

La estratificación dinámica de respuesta al tratamiento se evaluó en cuatro categorías: respuesta excelente, respuesta indeterminada, respuesta bioquímicamente incompleta y respuesta estructuralmente incompleta.

Como dominio secundario de validación, se evaluó la concordancia para el estadio AJCC.

Análisis estadístico

La herramienta no fue considerada un estándar de referencia clínico absoluto, sino un algoritmo preespecificado frente al cual se evaluó la concordancia de las clasificaciones asignadas por los expertos.

La concordancia exacta entre las respuestas de los expertos y las clasificaciones generadas por el algoritmo se evaluó mediante el porcentaje de acuerdo observado y el coeficiente kappa no ponderado. Dado el carácter ordinal de las categorías de riesgo inicial, respuesta dinámica y estadio AJCC, se calculó además el coeficiente kappa ponderado con ponderación lineal como análisis complementario, con el objetivo de considerar la magnitud de las discrepancias entre categorías. Para cada dominio se incluyeron únicamente las respuestas clasificables, por lo que las evaluaciones no homologables o ausentes fueron excluidas del análisis correspondiente.

Las categorías textuales consignadas por los

expertos fueron previamente homologadas con las categorías de salida del algoritmo. La homologación se realizó antes del cálculo de las medidas de concordancia y de acuerdo con reglas predefinidas para cada dominio.

Para la estratificación inicial del riesgo de recurrencia se evaluó, en primer lugar, la concordancia exacta en cuatro categorías: bajo, intermedio-bajo, intermedio-alto y alto. Como análisis complementario, se agruparon las categorías intermedio-bajo e intermedio-alto en una única categoría de riesgo intermedio y se recalculó la concordancia para una clasificación en tres categorías: bajo, intermedio y alto.

Para la estratificación dinámica de respuesta se evaluó la concordancia exacta entre las cuatro categorías definidas por el algoritmo: respuesta excelente, indeterminada, bioquímicamente incompleta y estructuralmente incompleta.

La concordancia para el estadio AJCC se analizó como dominio secundario.

Además del análisis global, se realizó un análisis descriptivo por experto y por caso. La concordancia por experto se expresó como el porcentaje de evaluaciones coincidentes con el algoritmo en cada dominio. El análisis por caso permitió identificar los escenarios clínicos con mayor y menor concentración de discrepancias entre los expertos y el algoritmo.

Los porcentajes de acuerdo observado y los coeficientes de concordancia se acompañaron de sus respectivos intervalos de confianza del 95%. Dado que cada caso fue evaluado por múltiples expertos, la incertidumbre de las estimaciones se evaluó mediante bootstrap no paramétrico por conglomerados a nivel del caso clínico, con 1000 remuestras, remuestreando los casos con reemplazo y conservando, en cada remuestra, todas las evaluaciones experto-caso correspondientes. Como análisis de sensibilidad frente a posibles efectos derivados de distribuciones marginales desequilibradas, se calculó además el coeficiente AC1 de Gwet.

Herramienta digital

La calculadora evaluada fue desarrollada como una aplicación web progresiva, con po-

sibilidad de ejecución en navegadores web e instalación en dispositivos móviles. Los algoritmos clínicos fueron implementados en JavaScript y organizados en módulos independientes correspondientes a la estratificación inicial del riesgo de recurrencia, la estadificación TNM/AJCC y la estratificación dinámica de respuesta al tratamiento.

La consolidación de las respuestas de los expertos, la homologación de las categorías y los análisis de concordancia se realizaron a partir de bases de datos estructuradas.

Los análisis estadísticos se realizaron con R, versión 4.2.3, mediante los paquetes readxl, dplyr, tidyr, ggplot2 e irr. El cálculo del AC1 de Gwet y los intervalos de confianza por bootstrap se realizó mediante funciones ad hoc programadas en R [9].

RESULTADOS

Características generales de la validación

Participaron 9 expertos con experiencia en CDT, quienes evaluaron de manera independiente 50 casos clínicos sintéticos contruidos como escenarios hipotéticos estandarizados. En conjunto, se obtuvieron 450 evaluaciones experto-caso (Tabla 1; Figura 1). Los casos abarcaron las categorías preespecificadas de riesgo inicial y respuesta dinámica y representaron el espectro clínico y anatomopatológico definido previamente en el diseño del estudio.

Concordancia en la estratificación inicial del riesgo de recurrencia

La concordancia global entre los expertos y el algoritmo para la estratificación inicial del riesgo en cuatro categorías fue del 88,4% (IC95%: 83,0–92,9), con un coeficiente kappa no ponderado de 0,840 (IC95%: 0,758–0,900), un kappa ponderado lineal de 0,844 (IC95%: 0,776–0,899) y un AC1 de Gwet de 0,859 (IC95%: 0,781–0,912) (Tabla 2). Los resultados globales de concordancia para los distintos dominios evaluados se resumen gráficamente en la Figura 2.

Al agrupar las categorías intermedio-bajo e intermedio-alto en una única categoría de riesgo intermedio, la concordancia aumentó al

91,8% (IC95%: 87,7–95,1), con un kappa no ponderado de 0,876 (IC95%: 0,809–0,926), un kappa ponderado lineal de 0,869 (IC95%: 0,805–0,923) y un AC1 de Gwet de 0,877 (IC95%: 0,817–0,931).

La matriz de concordancia para la clasificación agrupada mostró una marcada

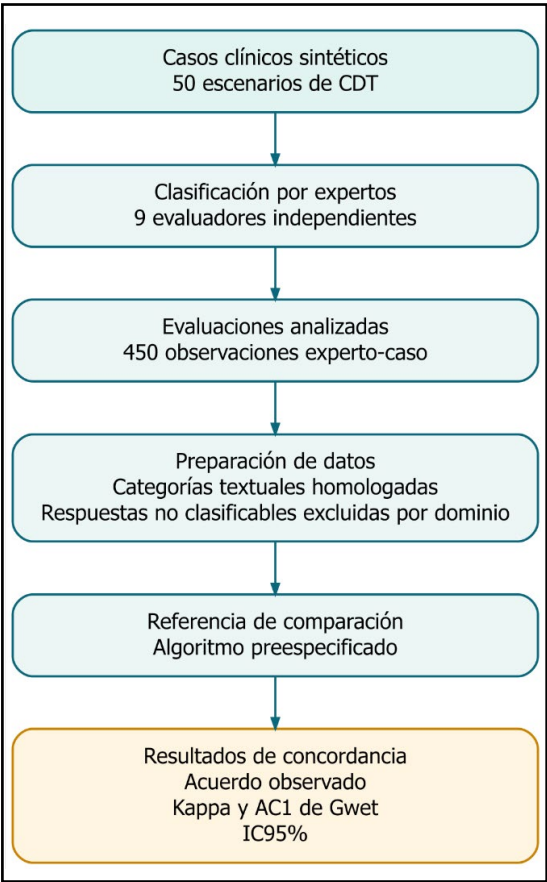


Figura 1. Flujo del estudio de validación. Se diseñaron 50 casos clínicos sintéticos, contruidos como escenarios hipotéticos estandarizados de carcinoma diferenciado de tiroides, y evaluados de manera independiente por 9 expertos. Las 450 evaluaciones experto-caso fueron comparadas con un algoritmo preespecificado luego de la homologación de categorías textuales y la exclusión de respuestas no clasificables por dominio. La concordancia se evaluó mediante acuerdo observado, kappa no ponderado y kappa ponderado lineal, con AC1 de Gwet como análisis de sensibilidad. Se estimaron intervalos de confianza del 95%.

Tabla 1. Diseño de la validación

Característica	Valor
Expertos participantes	9
Casos clínicos sintéticos	50
Evaluaciones experto-caso	450
Dominios principales de validación	Riesgo inicial de recurrencia; respuesta dinámica al tratamiento
Dominio secundario de validación	Estadio AJCC
Categorías de riesgo inicial	Bajo; intermedio-bajo; intermedio-alto; alto
Categorías de respuesta dinámica	Excelente; indeterminada; bioquímicamente incompleta; estructuralmente incompleta
Comparación evaluada	Clasificación de expertos frente a algoritmo clínico preespecificado

Cada uno de los 50 casos clínicos sintéticos, contruidos como escenarios hipotéticos estandarizados, fue evaluado de manera independiente por 9 expertos.

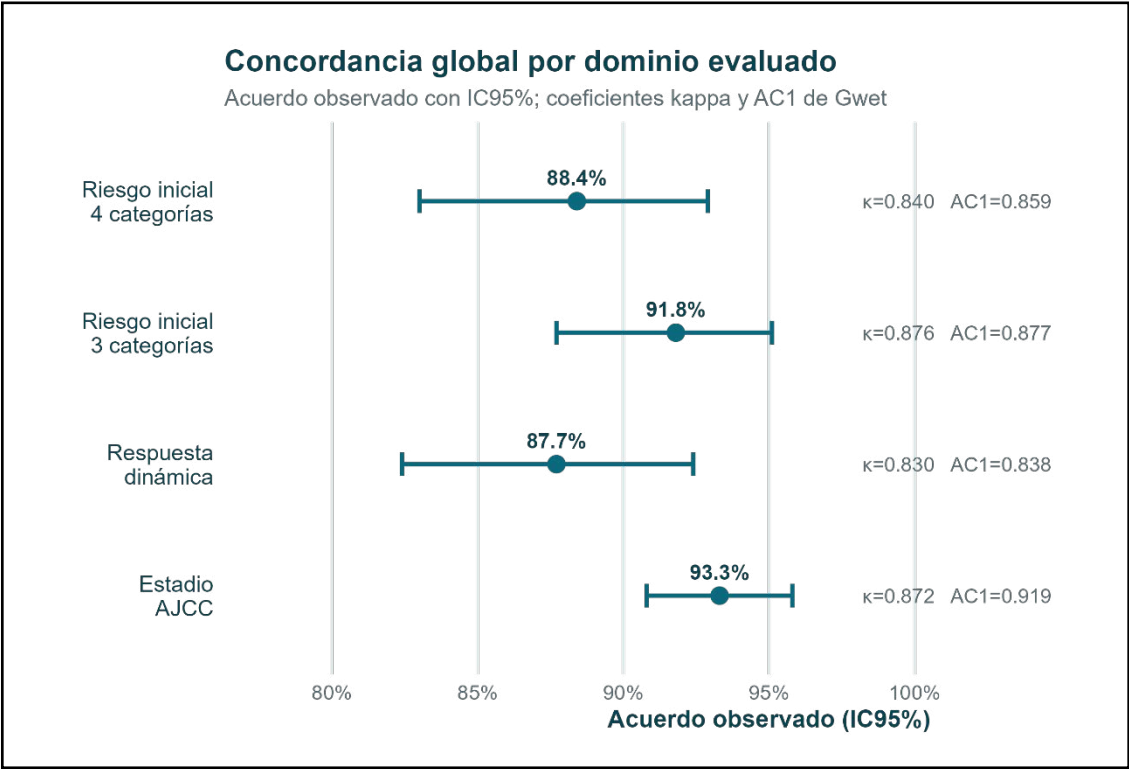


Figura 2. Concordancia global entre la clasificación de los expertos y el algoritmo preespecificado según dominio evaluado. Los puntos representan el acuerdo observado y las líneas horizontales sus intervalos de confianza del 95%, calculados mediante bootstrap no paramétrico por conglomerados a nivel del caso clínico. A la derecha se muestran el kappa no ponderado y el AC1 de Gwet. El kappa ponderado lineal se presenta en la Tabla 2.

concentración de observaciones en la diagonal principal (Tabla 3). No se observaron discrepancias directas entre las categorías de bajo y alto riesgo. Las discordancias se concentraron en transiciones entre categorías adyacentes: 20 evaluaciones entre riesgo bajo e intermedio y 17 entre riesgo intermedio y alto.

La concordancia por experto osciló entre 76% y 96% para la clasificación en cuatro categorías y entre 80% y 98% para la clasificación agrupada en tres categorías (Tabla 4).

Concordancia en la estratificación dinámica de respuesta al tratamiento

La concordancia global para la estratificación dinámica de respuesta fue del 87,7% (IC95%: 82,4–92,4), con un kappa no ponderado de 0,830 (IC95%: 0,747–0,896), un kappa ponderado lineal de 0,812 (IC95%: 0,715–0,887) y un AC1 de Gwet de 0,838 (IC95%: 0,770–0,900) (Tabla 2).

La distribución cruzada de las clasificaciones asignadas por los expertos y el algoritmo mostró una elevada concentración

Tabla 2. Concordancia global entre expertos y algoritmo en los dominios de validación

Dominio evaluado	n analizado	Acuerdo observado, % (IC95%)	Kappa no ponderado (IC95%)	Kappa ponderado lineal (IC95%)	AC1 de Gwet (IC95%)
Riesgo inicial, 4 categorías	449	88,4 (83,0–92,9)	0,840 (0,758–0,900)	0,844 (0,776–0,899)	0,859 (0,781–0,912)
Riesgo inicial, 3 categorías	449	91,8 (87,7–95,1)	0,876 (0,809–0,926)	0,869 (0,805–0,923)	0,877 (0,817–0,931)
Respuesta dinámica	448	87,7 (82,4–92,4)	0,830 (0,747–0,896)	0,812 (0,715–0,887)	0,838 (0,770–0,900)
Estadio AJCC	447	93,3 (90,8–95,8)	0,872 (0,825–0,916)	0,875 (0,834–0,911)	0,919 (0,884–0,945)

IC95%: intervalo de confianza del 95%. Los intervalos de confianza se estimaron mediante bootstrap no paramétrico por conglomerados a nivel del caso clínico, con 1000 remuestras, remuestreando los casos con reemplazo y conservando, en cada remuestra, todas las evaluaciones experto-caso correspondientes. Las respuestas ausentes o no homologables fueron excluidas del análisis del dominio correspondiente. El kappa ponderado lineal se calculó como análisis complementario para considerar la magnitud de las discrepancias entre categorías ordinales. El coeficiente AC1 de Gwet se calculó como análisis de sensibilidad.

Tabla 3. Matriz de concordancia para la estratificación inicial del riesgo de recurrencia en tres categorías

Clasificación del experto	Algoritmo: bajo	Algoritmo: intermedio	Algoritmo: alto	Total
Bajo	119	14	0	133
Intermedio	6	136	14	156
Alto	0	3	157	160
Total	125	153	171	449

Las celdas de la diagonal principal representan evaluaciones concordantes entre el experto y el algoritmo. Las categorías intermedio-bajo e intermedio-alto se agruparon en una única categoría de riesgo intermedio. No se observaron discrepancias directas entre las categorías de bajo y alto riesgo.

de observaciones en la diagonal principal (Tabla 5). Las discrepancias fueron más frecuentes entre categorías adyacentes. En particular, 19 evaluaciones mostraron discordancia entre respuesta indeterminada y bioquímicamente incompleta, mientras que 22 presentaron discrepancias entre respuesta excelente e indeterminada. Las discrepancias que involucraron la categoría de respuesta estructuralmente incompleta fueron menos frecuentes que las observadas entre las restantes categorías. La con-

cordancia por experto osciló entre 64% y 96% (Tabla 4).

Concordancia para el estadio AJCC

La concordancia global entre los expertos y el algoritmo para el estadio AJCC fue del 93,3% (IC95%: 90,8–95,8), con un kappa no ponderado de 0,872 (IC95%: 0,825–0,916), un kappa ponderado lineal de 0,875 (IC95%: 0,834–0,911) y un AC1 de Gwet de 0,919 (IC95%: 0,884–0,945) (Tabla 2). La concordancia por experto osciló entre 81,2% y 100% (Tabla 4).

Tabla 4. Concordancia entre expertos y algoritmo según evaluador.

Experto	n	Riesgo inicial, 4 categorías	Riesgo inicial, 3 categorías	Respuesta dinámica	Estadio AJCC
Experto 1	50	90%	92%	96%	100%
Experto 2	50	92%	94%	92%	87,8%
Experto 3	50	76%	80%	64%	86%
Experto 4	50	90%	96%	88%	94%
Experto 5	50	86%	90%	80%	94%
Experto 6	50	88%	88%	86%	96%
Experto 7	50	96%	98%	96%	81,2%
Experto 8	50	84%	88%	94%	100%
Experto 9	50	92%	98%	90%	100%

Los valores corresponden al porcentaje de acuerdo observado entre las clasificaciones de cada experto y las generadas por el algoritmo.

Tabla 5. Matriz de concordancia para la estratificación dinámica de respuesta al tratamiento

Clasificación del experto	Algoritmo: excelente	Algoritmo: indeterminada	Algoritmo: bioquímicamente incompleta	Algoritmo: estructuralmente incompleta	Total
Excelente	118	10	0	1	129
Indeterminada	12	137	4	1	154
Bioquímicamente incompleta	4	15	86	0	105
Estructuralmente incompleta	0	8	0	52	60
Total	134	170	90	54	448

Las celdas de la diagonal principal representan evaluaciones concordantes entre el experto y el algoritmo.

Análisis por experto y por caso

La concordancia individual mostró variabilidad entre evaluadores y entre dominios. Para la estratificación inicial del riesgo en cuatro categorías, los valores oscilaron entre 76% y 96%; para el riesgo agrupado, entre 80% y 98%; para la respuesta dinámica, entre 64% y 96%; y para el estadio AJCC, entre 81,2% y 100% (Tabla 4).

La mayor variabilidad entre expertos se observó en la estratificación dinámica de respuesta y en la clasificación del riesgo inicial en cuatro categorías. No se identificó un patrón uniforme de menor concordancia de un mismo evaluador en todos los dominios.

El análisis por caso mostró que las discrepancias no se distribuyeron de manera homogénea entre los 50 escenarios clínicos, sino que

se concentraron en un subconjunto de casos con mayor variabilidad entre los expertos y el algoritmo. Los casos con mayor discordancia correspondieron principalmente a escenarios de riesgo intermedio o respuesta no categórica, en los que coexistían factores modificadores como compromiso ganglionar limitado, márgenes microscópicos, invasión vascular, valores intermedios de tiroglobulina, anticuerpos anti-tiroglobulina positivos o hallazgos estructurales inespecíficos. La distribución de las concordancias y discrepancias para cada combinación experto-caso se presenta en la Figura 3.

DISCUSIÓN

En este estudio, un algoritmo clínico digital preespecificado mostró una elevada concordancia con la clasificación independiente rea-

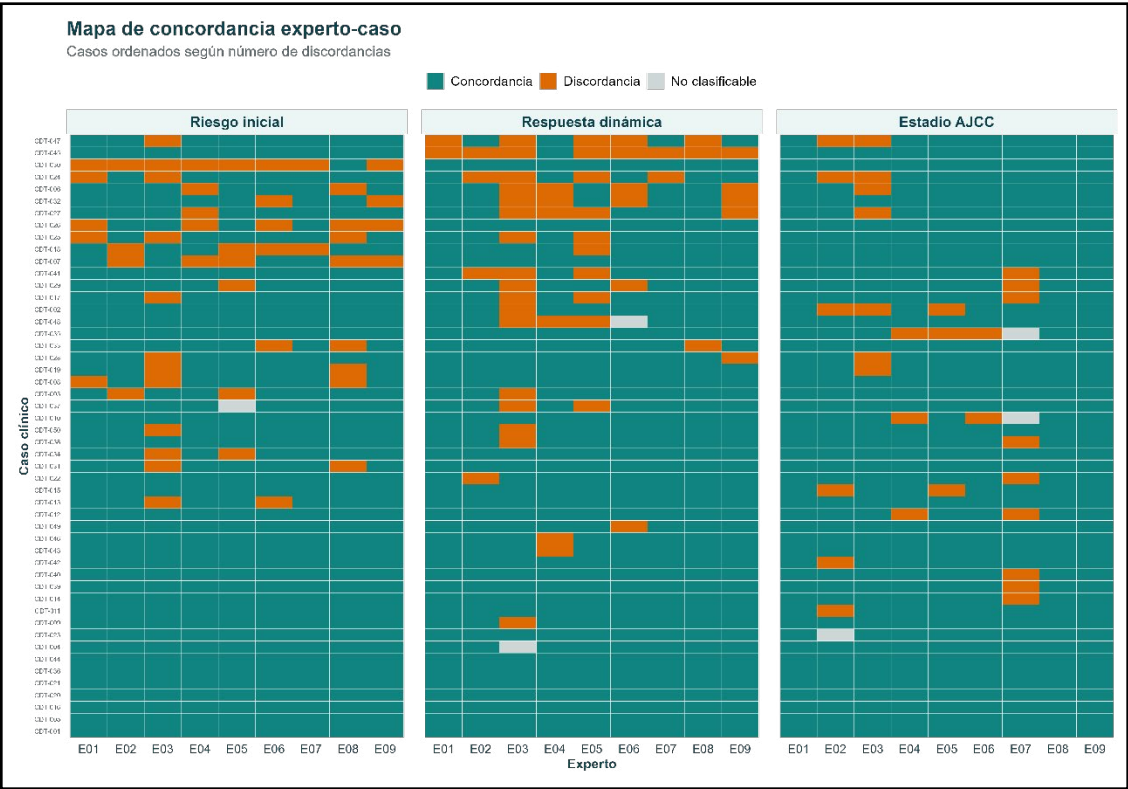


Figura 3. Mapa de concordancia experto-caso para riesgo inicial, estratificación dinámica y estadio AJCC. Cada celda representa una evaluación experto-caso. El verde indica concordancia entre el experto y el algoritmo preespecificado, el naranja discordancia y el gris respuestas no clasificables. Los casos fueron ordenados según el número total de discordancias para facilitar la identificación de escenarios con mayor variabilidad.

lizada por expertos en los dominios principales de validación: la estratificación inicial del riesgo de recurrencia y la evaluación dinámica de la respuesta al tratamiento, así como en el estadio AJCC. La consistencia observada mediante diferentes coeficientes de concordancia respalda la robustez del hallazgo y sugiere que criterios clínicos complejos pueden ser trasladados a una herramienta digital estructurada manteniendo una adecuada correspondencia con el juicio experto [1,3,7,9,10].

Un hallazgo particularmente relevante fue que las discordancias no se distribuyeron aleatoriamente, sino que se concentraron predominantemente entre categorías adyacentes. En la estratificación inicial del riesgo no se observaron discrepancias directas entre los extremos de bajo y alto riesgo. Al agrupar las categorías intermedio-bajo e intermedio-alto en una única categoría de riesgo intermedio, el acuerdo aumentó de 88,4% a 91,8%, mientras que el kappa no ponderado pasó de 0,840 a 0,876. Los valores elevados del kappa ponderado lineal en ambas clasificaciones –0,844 para cuatro categorías y 0,869 para tres categorías– confirmaron asimismo que la magnitud de las discrepancias fue, en general, limitada. Estos resultados sugieren que una proporción relevante de las diferencias se origina en la subdivisión fina del riesgo intermedio y no en clasificaciones extremas o diametralmente opuestas.

Un patrón semejante se observó en la estratificación dinámica de respuesta. Las discordancias se concentraron principalmente entre respuesta excelente e indeterminada y entre respuesta indeterminada y bioquímicamente incompleta. Las discrepancias que involucraron la respuesta estructuralmente incompleta fueron menos frecuentes y se concentraron fundamentalmente en su distinción respecto de la respuesta indeterminada. El kappa ponderado lineal fue de 0,812, manteniéndose en un rango elevado y respaldando la interpretación de que, cuando existieron diferencias entre el algoritmo y los expertos, estas ocurrieron principalmente entre categorías relativamente próximas dentro de la clasificación.

La concordancia fue numéricamente mayor para el estadio AJCC que para la estratificación dinámica de respuesta, con acuerdos observados de 93,3% y 87,7%, respectivamente. Una posible explicación es la diferente naturaleza de ambos sistemas. La estratificación AJCC está orientada principalmente a la estimación de mortalidad específica y se construye a partir de un conjunto relativamente acotado de variables predefinidas, mientras que la estratificación dinámica requiere integrar información bioquímica y estructural obtenida durante el seguimiento, incluidos los valores de tiroglobulina, la evolución de los anticuerpos antitiroglobulina y los hallazgos en los estudios por imágenes [1-4]. Esta mayor complejidad integrativa podría contribuir a la mayor dispersión observada entre expertos en este dominio, en el cual la concordancia individual osciló entre 64% y 96%. En contraste, para el estadio AJCC la concordancia por experto se mantuvo entre 81,2% y 100%.

La variabilidad entre evaluadores fue diferente según el dominio analizado. No se identificó un patrón uniforme de menor concordancia de un mismo experto en todas las clasificaciones, lo que sugiere que las discrepancias fueron, al menos parcialmente, dependientes de la tarea clínica evaluada y no exclusivamente de diferencias generales entre los participantes. Esta observación resulta relevante porque ilustra la heterogeneidad que puede surgir en la aplicación manual de criterios complejos y multicomponente, incluso cuando los evaluadores analizan los mismos casos estandarizados y comparten un marco conceptual común. En este contexto, una herramienta algorítmica preespecificada podría contribuir a una aplicación más homogénea de las reglas de clasificación; sin embargo, este posible efecto no fue evaluado directamente en el presente estudio y deberá analizarse mediante diseños específicos que comparen la práctica con y sin apoyo de la herramienta.

La distribución de las discrepancias por caso aportó una dimensión adicional al análisis. Las diferencias no se distribuyeron de manera uniforme entre los 50 escenarios clínicos, sino que

se concentraron en un subconjunto de casos con mayor variabilidad entre los expertos y el algoritmo. Este hallazgo sugiere la existencia de escenarios particularmente sensibles a diferencias de interpretación o a la aplicación de reglas de combinación. La caracterización específica de estos casos podría ser útil en estudios posteriores para identificar áreas de mayor incertidumbre clasificatoria y optimizar tanto la implementación del algoritmo como las estrategias educativas vinculadas con el uso de los sistemas de estratificación.

Características diferenciales de la herramienta evaluada

La herramienta evaluada comparte con otras calculadoras digitales recientes, como la *ATA 2025 Risk of Recurrence Calculator* de TIRO [8], el objetivo de trasladar los criterios contemporáneos de estratificación del CDT a una interfaz clínica estructurada. Sus principales características diferenciales se relacionan con la integración de la estadificación pronóstica AJCC, la lógica aplicada a la estratificación dinámica y su adaptación al contexto clínico regional.

Una primera característica diferencial es la incorporación explícita de la estadificación pronóstica AJCC/UICC 8.^a edición. La herramienta calcula no solo los componentes T, N y M, sino también el estadio final, integrando automáticamente la edad del paciente, variable determinante para la agrupación por estadios en el CDT. De este modo, dentro de un mismo flujo clínico se vinculan dos sistemas con objetivos diferentes y complementarios: la estratificación del riesgo de persistencia o recurrencia y la estimación pronóstica de mortalidad específica por cáncer [1,3]. A diferencia de la calculadora ATA 2025 de TIRO, centrada en la estimación del riesgo de recurrencia, la herramienta evaluada incorpora, además, el cálculo automatizado del estadio AJCC [8].

Una segunda característica diferencial se relaciona con la lógica empleada para la estratificación dinámica. La herramienta integra de manera conjunta los valores absolutos de

tiroglobulina, los anticuerpos antitiroglobulina y los hallazgos de los estudios por imágenes. En particular, evita que una tendencia aislada de tiroglobulina determine automáticamente un cambio de categoría cuando los valores absolutos permanecen dentro del rango con una respuesta excelente. Esta enfoque procura preservar una evaluación integrada de la información bioquímica y estructural, evitando reclasificaciones desproporcionadas en pacientes con niveles bioquímicos muy bajos. El impacto de esta estrategia sobre la clasificación de los pacientes y las decisiones clínicas deberá ser evaluado específicamente en futuros estudios de implementación.

Finalmente, la herramienta fue desarrollada en español, adaptada al contexto clínico local y alineada con el Consenso Delphi Intersocietario argentino sobre diagnóstico, tratamiento y seguimiento del CDT [7]. Esta contextualización puede favorecer su aplicabilidad en la práctica endocrinológica regional y facilitar la implementación estructurada de recomendaciones internacionales y nacionales dentro de un mismo proceso de evaluación clínica.

Fortalezas y limitaciones

Entre las principales fortalezas del estudio se encuentra la evaluación independiente y ciega respecto de las clasificaciones generadas por el algoritmo y de las respuestas de los demás participantes. El uso de casos estandarizados permitió además que todos los expertos fueran expuestos a los mismos escenarios clínicos, reduciendo la variabilidad atribuible a diferencias en la información disponible.

Desde el punto de vista estadístico, la concordancia se evaluó mediante varios enfoques complementarios. El porcentaje de acuerdo observado y el kappa no ponderado permitieron cuantificar la coincidencia exacta entre las clasificaciones. Dado el carácter ordinal de las clasificaciones evaluadas, el kappa ponderado lineal permitió considerar adicionalmente la magnitud de las discrepancias entre categorías. Asimismo, el AC1 de Gwet se utilizó como análisis de sensibilidad frente a posi-

bles efectos derivados de distribuciones marginales desequilibradas [9]. La convergencia de las estimaciones obtenidas mediante estos diferentes coeficientes refuerza la robustez de los resultados.

Los intervalos de confianza se estimaron mediante bootstrap no paramétrico por conglomerados a nivel del caso clínico, con 1000 remuestras, preservando en cada remuestra todas las evaluaciones experto-caso correspondientes. Este procedimiento permitió respetar la dependencia existente entre las múltiples evaluaciones realizadas sobre un mismo escenario clínico.

El estudio presenta, sin embargo, varias limitaciones. En primer lugar, se utilizaron casos clínicos sintéticos, contruidos como escenarios hipotéticos estandarizados y generados específicamente para el estudio. Estos casos fueron diseñados deliberadamente para cubrir un espectro amplio de categorías de riesgo y respuesta. Esta estrategia fortalece la evaluación metodológica del algoritmo, pero no reproduce necesariamente la prevalencia relativa de las distintas categorías ni la distribución de la complejidad clínica observada en la práctica real.

En segundo lugar, la concordancia no equivale a exactitud pronóstica. En ausencia de desenlaces clínicos de referencia, como persistencia, recurrencia o mortalidad específica, el presente estudio no permite establecer si las clasificaciones generadas por el algoritmo, o las asignadas por los expertos, predicen correctamente la evolución de los pacientes. Los resultados demuestran concordancia entre métodos de clasificación, pero no validez predictiva frente a desenlaces clínicos reales.

En tercer lugar, el número de expertos fue limitado y su selección fue intencional. Todos habían participado en el Consenso Delphi Intersocietario argentino y contaban con experiencia en el manejo del CDT, lo que fortalece la pertinencia del panel para el contexto del estudio, pero puede limitar la generalización de los resultados a profesionales con otros niveles de experiencia, ámbitos asistenciales o marcos de referencia.

En cuarto lugar, los participantes pertenecían a un mismo contexto nacional y compartían exposición al Consenso Delphi argentino, lo que podría haber favorecido cierta convergencia conceptual. La reproducibilidad de estos resultados deberá evaluarse en grupos de expertos de otros países y en contextos donde las decisiones se apoyen exclusivamente en recomendaciones internacionales.

En quinto lugar, el estudio no tuvo como objetivo principal evaluar la concordancia interexperto ni comparar directamente la variabilidad de las decisiones con y sin apoyo del algoritmo. Por lo tanto, aunque la estandarización de reglas podría contribuir a una aplicación más homogénea de los criterios, no puede concluirse a partir de estos datos que la herramienta reduzca efectivamente la variabilidad entre profesionales.

Por último, la evaluación se centró en la concordancia clasificatoria y no incluyó variables de implementación, como tiempo requerido para la toma de decisiones, facilidad de uso, aceptabilidad por parte de los profesionales o impacto sobre la conducta clínica. Estos aspectos serán necesarios para determinar la utilidad de la herramienta en la práctica asistencial.

Los hallazgos del presente estudio proporcionan una base metodológica para etapas posteriores de validación. Entre ellas se incluyen la evaluación prospectiva de la herramienta en pacientes reales, el análisis de su desempeño frente a desenlaces objetivos de persistencia y recurrencia, la validación en poblaciones y contextos geográficos diferentes y los estudios de implementación destinados a determinar su impacto sobre el tiempo de decisión, la reproducibilidad de las clasificaciones y la variabilidad entre profesionales.

En conclusión, el algoritmo digital evaluado mostró una elevada concordancia con las clasificaciones asignadas por expertos para la estratificación inicial del riesgo de recurrencia, la respuesta dinámica al tratamiento y el estadio AJCC. Los resultados fueron consistentes mediante kappa no ponderado, kappa ponderado lineal y AC1 de Gwet, y las discor-

dancias se concentraron predominantemente entre categorías adyacentes. Estos hallazgos respaldan el potencial de un algoritmo estructurado como herramienta de apoyo a la clasificación pronóstica en CDT, sujeto a validación prospectiva y estudios de implementación en la práctica real.

Conflicto de intereses

LARS participó en el diseño y desarrollo de la herramienta digital evaluada en este estudio. El desarrollo de la herramienta no recibió financiación comercial. Los demás autores declaran no tener conflictos de interés relacionados con este trabajo.

Contribuciones de los autores

LARS fue responsable de la conceptualización del estudio, el diseño general del trabajo, la coordinación del proyecto y la redacción del borrador inicial del manuscrito. Todos los autores contribuyeron a la revisión crítica del contenido intelectual, realizaron correcciones y aportes relevantes, aprobaron la versión final para su publicación y aceptan la responsabilidad sobre el contenido del trabajo.

Financiación

El presente estudio y el desarrollo de la herramienta digital no recibieron financiación comercial ni financiación externa específica.

Consideraciones éticas

El estudio utilizó exclusivamente casos clínicos sintéticos, contruidos como escenarios hipotéticos estandarizados, y no incluyó datos de pacientes reales ni información identificable. La participación de los expertos fue voluntaria. No se recopilaron datos personales sensibles de los evaluadores.

Disponibilidad de datos y código

Los datos anonimizados derivados de las evaluaciones de los expertos y el código utilizado para los análisis estadísticos estarán disponibles previa solicitud razonable al autor de correspondencia. Las reglas clínicas implementadas por el algoritmo se describen en el manuscrito y derivan de las fuentes citadas. El código fuente de la aplicación web no es de acceso público.

REFERENCIAS

1. Ringel MD, Sosa JA, Baloch Z, Bischoff L, Bloom G, Brent GA, et al. 2025 American Thyroid Association Management Guidelines for Adult Patients with Differentiated Thyroid Cancer. *Thyroid*. 2025; 35(8):841-985. doi: 10.1177/10507256251363120.
2. Tuttle RM, Alzahrani AS. Risk Stratification in Differentiated Thyroid Cancer: From Detection to Final Follow-Up. *J Clin Endocrinol Metab*. 2019; 104(9):4087-100. doi: 10.1210/jc.2019-00177.
3. Tuttle RM, Haugen B, Perrier ND. Updated American Joint Committee on Cancer/Tumor-Node-Metastasis Staging System for Differentiated and Anaplastic Thyroid Cancer (Eighth Edition): What Changed and Why? *Thyroid*. 2017; 27(6):751-6. doi: 10.1089/thy.2017.0102.
4. Tuttle RM, Tala H, Shah J, Leboeuf R, Ghossein R, Gonen M, et al. Estimating Risk of Recurrence in Differentiated Thyroid Cancer After Total Thyroidectomy and Radioactive Iodine Remnant Ablation: Using Response to Therapy Variables to Modify the Initial Risk Estimates Predicted by the New American Thyroid Association Staging System. *Thyroid*. 2010; 20(12):1341-1349. doi: 10.1089/thy.2010.0178.
5. Momesso DP, Vaisman F, Yang SP, Bulzico DA, Corbo R, Vaisman M, Tuttle RM. Dynamic Risk Stratification in Patients with Differentiated Thyroid Cancer Treated Without Radioactive Iodine. *J Clin Endocrinol Metab*. 2016; 101(7):2692-700. doi: 10.1210/jc.2015-4290.
6. Park S, Kim WG, Song E, Oh HS, Kim M, Kwon H, et al. Dynamic Risk Stratification

- for Predicting Recurrence in Patients with Differentiated Thyroid Cancer Treated Without Radioactive Iodine Remnant Ablation Therapy. *Thyroid*. 2017; 27(4):524-530. doi: 10.1089/thy.2016.0477.
7. Pitoia F, Abelleira E, Galíndez MM, Poenitz F, Melado GE, Falco A, et al. Consenso Delphi intersocietario argentino sobre diagnóstico, tratamiento y seguimiento del carcinoma diferenciado de tiroides: FASEN | SAEM | AACCyC | AABYMN | AAOC. *EndoAr*. 2026; 2(2):92-130. doi: 10.65640/4c6m9685.
8. TIRO. ATA 2025 Risk of Recurrence Calculator [Internet]. tiro.expert. Consultado julio 2026. Disponible en: <https://tiro.expert/ata-2025-risk-of-recurrence-calculator/>
9. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol*. 2008; 61(Pt 1):29-48. doi: 10.1348/000711006X126600.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33(1):159-174.