

**Inteligencia artificial en la detección temprana del cáncer de laringe mediante
análisis de imágenes endoscópicas**

Artificial intelligence in the early detection of laryngeal cancer through endoscopic
image analysis

Inteligência artificial na detecção precoce do câncer de laringe por meio de análise de
imagens endoscópicas.

Allister Antonio Miñán Vera
Instituto Europeo de Postgrados
Ecuador
allister1999@hotmail.com
<https://orcid.org/0009-0006-3916-6739>

Forma de citación en APA, séptima edición.

Miñán, A. (2026). *Inteligencia artificial en la detección temprana del cáncer de laringe mediante análisis de imágenes endoscópicas*. Revista Ibero Research, 1(7), 1110 - 1144.

Fecha de presentación: 20/09/2026

Fecha de aceptación: 04/10/2026

Fecha de publicación: 06/10/2026

Resumen

La inteligencia artificial aplicada a endoscopia laríngea muestra potencial para apoyar detección, clasificación y segmentación de lesiones. Se realizó una revisión narrativa crítica de estudios, metaanálisis y estándares de reporte. Se diferenciaron pacientes e imágenes, tareas diagnósticas y tipos de validación. Los resultados publicados muestran desempeño variable y experiencias de validación externa y vídeo, pero no acreditan por sí solos diagnóstico más temprano o mejores desenlaces oncológicos. Se propone una matriz de evaluación e implementación supervisada que integra calidad de datos, seguridad e impacto clínico. Se concluye que la adopción requiere validación independiente, control de sesgos y seguimiento de decisiones, manteniendo confirmación diagnóstica y responsabilidad especializada.

Palabras clave: inteligencia artificial; cáncer de laringe; endoscopia; aprendizaje profundo; exactitud diagnóstica.

Abstract

Artificial intelligence applied to laryngeal endoscopy may support lesion detection, classification, and segmentation. A critical narrative review examined studies, meta-analyses, and reporting standards. Patients and images, diagnostic tasks, and validation types were differentiated. Published results show variable performance and experience with external and video validation, but do not independently establish earlier diagnosis or improved oncological outcomes. An evaluation and supervised implementation matrix integrating data quality, safety, and clinical impact is proposed. Adoption requires independent validation, bias control, and monitoring of decisions while maintaining diagnostic confirmation and specialist responsibility.

Keywords: artificial intelligence; laryngeal cancer; endoscopy; deep learning; diagnostic accuracy.

Resumo

A inteligência artificial aplicada à endoscopia laríngea apresenta potencial para apoiar detecção, classificação e segmentação de lesões. Realizou-se uma revisão narrativa crítica de estudos, metanálises e padrões de relato. Diferenciaram-se pacientes e imagens, tarefas diagnósticas e tipos de validação. Os resultados publicados mostram desempenho variável e experiências de validação externa e por vídeo, mas não demonstram, isoladamente, diagnóstico mais precoce ou melhores resultados oncológicos. Propõe-se uma matriz de avaliação e implementação supervisionada que integra qualidade dos dados, segurança e impacto clínico. A adoção exige validação independente, controle de vieses e acompanhamento das decisões, mantendo confirmação diagnóstica e responsabilidade especializada.

Palavras-chave: inteligência artificial; câncer de laringe; endoscopia; aprendizagem profunda; acurácia diagnóstica.

Introducción

El cáncer de laringe representa una enfermedad de considerable importancia dentro de la oncología de cabeza y cuello. La mayoría de los tumores laríngeos corresponden a carcinomas escamocelulares, que pueden desarrollarse en regiones glóticas, supraglóticas o subglóticas. Su aparición se encuentra relacionada con múltiples factores, entre ellos el consumo de tabaco, alcohol y determinadas exposiciones ambientales y ocupacionales. Además, pueden existir lesiones precursoras como leucoplasias, displasias y alteraciones epiteliales que requieren una valoración especializada debido a su potencial de progresión maligna (1,2).

La detección precoz constituye un componente esencial del abordaje del cáncer laríngeo. Las lesiones diagnosticadas en fases iniciales pueden presentar mejores posibilidades de tratamiento conservador y de preservación funcional. Sin embargo, la identificación temprana no siempre resulta sencilla, debido a que determinadas lesiones pueden presentar cambios sutiles de color, textura, vascularización o arquitectura superficial. La exploración endoscópica convencional mediante luz blanca continúa siendo una herramienta fundamental, pero su interpretación depende de factores relacionados con la calidad de la imagen, las características de la lesión y la experiencia del especialista. (1)

Paralelamente, el desarrollo de la inteligencia artificial ha modificado el panorama de la medicina basada en imágenes. El aprendizaje automático y, especialmente, el aprendizaje profundo permiten construir modelos capaces de identificar patrones complejos en imágenes digitales que pueden no ser fácilmente reconocibles mediante una inspección convencional. Las redes neuronales convolucionales (CNN) han adquirido particular importancia en la interpretación automatizada de imágenes médicas

debido a su capacidad para aprender representaciones jerárquicas de características visuales (3,4).

En laringología, la investigación sobre IA ha evolucionado desde modelos experimentales de clasificación de imágenes hacia sistemas capaces de detectar lesiones, diferenciar tejido benigno y maligno, identificar lesiones precursoras y delimitar los bordes tumorales. El propósito de esta revisión es analizar críticamente el papel potencial de la IA en la detección temprana del cáncer de laringe mediante imágenes endoscópicas y discutir los principales retos que deben resolverse antes de su implementación clínica generalizada. (1)

La detección temprana constituye una finalidad clínica más exigente que clasificar fotografías correctamente. Para demostrarla es necesario establecer si una herramienta adelanta el diagnóstico de lesiones relevantes, modifica decisiones y mejora resultados sin aumentar de forma desproporcionada biopsias o falsas alarmas. Los estudios de exactitud aportan un paso inicial, pero no responden por sí solos a esas preguntas asistenciales. (1,2,5)

Las tareas computacionales también deben diferenciarse. Detectar una lesión, clasificar su naturaleza y delimitar su superficie producen resultados distintos y utilizan métricas diferentes; una elevada concordancia de segmentación no representa sensibilidad diagnóstica ni capacidad para identificar invasión profunda. El artículo delimita su objeto al análisis de imágenes endoscópicas laríngeas, utilizando otras modalidades únicamente para discutir límites y complementariedad. (6,7,8)

La endoscopia proporciona información visual condicionada por iluminación, óptica, movimiento y selección de fotogramas. El rendimiento puede cambiar cuando un modelo entrenado con imágenes limpias se aplica a exploraciones con secreciones,

sangrado o visualización incompleta. La validación externa adquiere especial relevancia porque refleja, aunque de manera imperfecta, variación entre centros y equipos. (9,7)

Se plantea identificar qué evidencia respalda el apoyo diagnóstico, qué errores metodológicos pueden exagerarlo y qué condiciones debería cumplir una implementación supervisada. El aporte consiste en una matriz de evaluación que une calidad de datos, desempeño, interacción clínica y seguimiento; no se presenta un algoritmo nuevo ni resultados de entrenamiento propios.

Metodología

Se realizó una revisión narrativa crítica, de alcance documental y orientación aplicada.

El material temático inicial se amplió mediante búsquedas dirigidas en registros académicos y publicaciones originales disponibles; se verificaron identidad bibliográfica y correspondencia de las fuentes con las afirmaciones. La búsqueda complementaria se cerró el 5 de octubre de 2026 y conservó trabajos fundacionales cuando eran necesarios para interpretar la evidencia reciente.

Se consultaron registros de Europe PMC/PubMed, metadatos DOI y páginas de revistas u organismos pertinentes. Se incluyeron publicaciones con relación directa con la pregunta y datos o argumentos interpretables; se excluyeron duplicados bibliográficos, referencias que no correspondían al contenido atribuido y trabajos ajenos al objeto. No se declararon búsquedas en bases de suscripción no consultadas ni un proceso exhaustivo de revisión sistemática.

La selección fue intencional y se utilizó una matriz de extracción para ordenar diseño, población u objeto, resultados, medida y limitaciones. El análisis fue comparativo y temático; los resultados cuantitativos se conservaron en su contexto y las propuestas se diferenciaron de hallazgos empíricos. Las tablas y gráficos se elaboraron con

información publicada, sin crear participantes, observaciones o estimaciones de efecto propias.

La revisión utilizó fuentes documentales públicas y no intervino sobre personas ni accedió a registros individuales identificables. No se atribuye una aprobación ética inexistente. Se mantuvo trazabilidad mediante DOI o URL y se realizó una comprobación de correspondencia entre citas y referencias, junto con revisión de coherencia de cifras y elementos visuales.

Se realizó una revisión narrativa crítica con búsqueda complementaria de laryngeal, laryngoscopy, leukoplakia, artificial intelligence, deep learning y machine learning. Se priorizaron estudios con endoscopia laríngea, validación independiente o análisis de utilidad clínica; se incorporaron guías de reporte para interpretar aspectos de diseño. Los trabajos sobre voz, tomografía o histopatología no se utilizaron para atribuir desempeño a una modalidad endoscópica diferente.

La extracción identificó población, número de pacientes e imágenes, modalidad óptica, tarea, referencia diagnóstica y tipo de validación. Se registró la unidad de análisis porque miles de fotogramas de pocos pacientes no proporcionan la misma información que miles de personas independientes. Se distinguieron validación interna, externa y evaluación prospectiva, sin asumir que cualquiera de ellas implica beneficio asistencial demostrado.

Se interpretaron sensibilidad, especificidad, área bajo la curva y métricas de segmentación según su propósito. Los resultados de distintas revisiones no se combinaron nuevamente, dado el probable solapamiento de estudios y diferencias en selección y referencia. La comparación fue crítica y descriptiva; no se calculó una estimación única del rendimiento de toda la inteligencia artificial laríngea.

Las guías STARD y STARD-AI orientaron el examen de exactitud diagnóstica; TRIPOD+AI, la transparencia de modelos predictivos; CLAIM, la información específica de imágenes; y DECIDE-AI, la evaluación inicial en el flujo clínico. Se emplearon como marcos de lectura y no se declaró una puntuación formal de cumplimiento no realizada. (10,11,12,13,14)

Los gráficos conservan la tarea y el conjunto de evaluación de cada fuente. Los porcentajes de sensibilidad y especificidad se presentan como estimaciones publicadas, sin construir matrices de pacientes hipotéticos ni calcular valores predictivos sin prevalencia definida. La propuesta de implementación corresponde a una derivación metodológica de los hallazgos, no a una intervención probada.

Resultados

La laringoscopia permite examinar directamente las estructuras laríngeas y constituye uno de los principales procedimientos para la evaluación de pacientes con disfonía persistente, lesiones sospechosas u otros signos compatibles con patología laríngea. La visualización mediante luz blanca puede identificar irregularidades de la mucosa, masas, ulceraciones, cambios de color y alteraciones de la superficie. (15)

Sin embargo, existe una importante variabilidad interobservador. Algunas lesiones benignas pueden presentar características similares a lesiones malignas, mientras que lesiones premalignas o carcinomas incipientes pueden presentar manifestaciones discretas. La experiencia del otorrinolaringólogo influye directamente en la capacidad para reconocer patrones sospechosos. (15)

Estos resultados muestran que las imágenes endoscópicas contienen información morfológica y vascular potencialmente aprovechable por sistemas computacionales. La IA puede procesar grandes cantidades de imágenes y aprender asociaciones entre

características visuales y diagnósticos previamente establecidos mediante histopatología. (15)

La IA comprende diferentes técnicas computacionales destinadas a realizar tareas que tradicionalmente requieren capacidades cognitivas humanas. En medicina, el aprendizaje automático permite identificar patrones a partir de datos previamente etiquetados, mientras que el aprendizaje profundo emplea arquitecturas neuronales con múltiples capas capaces de generar representaciones complejas de los datos. (7)

Las CNN se han convertido en una de las principales herramientas para el análisis de imágenes médicas. Su arquitectura permite detectar progresivamente características como bordes, texturas, patrones geométricos y estructuras más complejas. En el caso de las imágenes laríngeas, estos modelos pueden aprender diferencias entre mucosa normal, lesiones benignas, lesiones premalignas y cáncer. (7)

Uno de los trabajos iniciales de gran relevancia fue desarrollado por Xiong y colaboradores (7), quienes utilizaron 13.721 imágenes laringoscópicas para construir y evaluar una red neuronal convolucional destinada a identificar cáncer laríngeo, lesiones precancerosas, tumores benignos y tejido normal. En un conjunto independiente procedente de otros hospitales, el modelo alcanzó una sensibilidad de 73,1%, especificidad de 92,2%, AUC de 0,922 y exactitud global de 86,7%.

Este trabajo fue importante porque demostró que el análisis computacional podía extenderse más allá de la clasificación binaria simple. El modelo intentaba diferenciar múltiples categorías de tejido, acercándose más a la complejidad del diagnóstico real. (7)

La tabla 1 organiza tareas de IA endoscópica y desenlaces, con indicación del alcance de cada componente y de las fuentes que sustentan su interpretación.

Tabla 1*Tareas de IA endoscópica y desenlaces*

Tarea	Resultado	Límite
Clasificación	Categoría de lesión	No sustituye confirmación (7)
Detección	Localización de área sospechosa	Depende de adquisición (9)
Segmentación	Delimitación superficial	No informa margen profundo (6)
Predicción de transformación	Riesgo futuro	Necesita seguimiento temporal (16)

Fuente: Síntesis de las publicaciones citadas.

Posteriormente, las investigaciones han incorporado algoritmos capaces de analizar secuencias de video, en lugar de limitarse a imágenes individuales. Wellenstein y colaboradores (9) desarrollaron un sistema de aprendizaje profundo capaz de localizar y clasificar carcinoma de cuerda vocal y lesiones benignas durante la endoscopia. El sistema alcanzó velocidades de procesamiento compatibles con aplicaciones potencialmente en tiempo real, lo que abre la posibilidad de que la IA funcione durante el procedimiento endoscópico.

Una revisión sistemática de Yao y colaboradores (8) identificó múltiples aplicaciones de IA en la laringoscopia ambulatoria, incluyendo clasificación automática de imágenes, detección de estructuras anatómicas y análisis de calidad de las imágenes. Este desarrollo resulta especialmente importante porque una herramienta de IA clínicamente útil no debería limitarse a emitir un diagnóstico final, sino también contribuir a garantizar que las imágenes obtenidas sean suficientemente informativas.

La evidencia acumulada durante los últimos años muestra resultados considerablemente prometedores. Żurek y colaboradores (17), en una revisión sistemática y metaanálisis sobre IA aplicada a la endoscopia laríngea, analizaron estudios que evaluaban imágenes obtenidas durante laringoscopia. Los modelos presentaron valores de exactitud entre 0,806 y 0,997. Para la diferenciación entre lesiones benignas y malignas, la sensibilidad y especificidad agrupadas fueron aproximadamente de 91% y 94%, respectivamente.

Estos resultados deben interpretarse con cautela. Un alto rendimiento en conjuntos de datos experimentales no significa automáticamente que un algoritmo pueda sustituir al especialista en condiciones clínicas reales. Las bases de datos pueden presentar imágenes de calidad superior a las obtenidas habitualmente en consulta y pueden incluir una distribución de enfermedades diferente de la encontrada en la población general. (17)

Xiong y colaboradores utilizaron 13.721 imágenes para desarrollar y evaluar clasificación de lesiones laríngeas. En la evaluación independiente de otros hospitales se reportaron sensibilidad de 73,1%, especificidad de 92,2%, área bajo la curva de 0,922 y exactitud de 86,7%. Estas métricas responden a contrastes definidos en el estudio; su coexistencia demuestra que una cifra de exactitud puede ocultar diferencias clínicamente importantes en errores. (7)

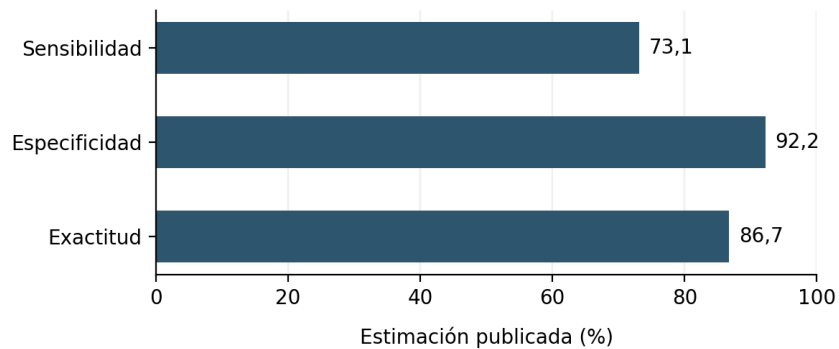
El rendimiento de la imagen asistida depende del sustrato óptico. La investigación de Ni y colaboradores sobre NBI muestra el valor de examinar patrones vasculares, pero la interpretación sigue vinculada a lesión y calidad de adquisición. La IA aprende asociaciones presentes en los datos; no convierte información ausente o una exploración incompleta en certeza histológica. (15,3,4)

La revisión de aplicaciones en laringoscopia ambulatoria describe funciones más amplias que clasificar cáncer. Reconocimiento anatómico, selección y calidad de imágenes pueden apoyar el procedimiento, aunque cada función necesita un desenlace propio. Un sistema útil para enseñanza o adquisición no debe promocionarse automáticamente como sustituto de biopsia o de juicio especializado. (8,18)

El gráfico 1 presenta rendimiento externo del modelo de Xiong. Se conserva la unidad y el contexto de los datos para distinguir la descripción publicada de las implicaciones propuestas.

Gráfico 1

Rendimiento externo del modelo de Xiong



Fuente: elaboración propia. (7). Evaluación independiente; métricas distintas, no categorías excluyentes.

La sensibilidad menor que la especificidad señala la necesidad de valorar falsos negativos, aun con exactitud global elevada. La comparación entre métricas no representa un contraste estadístico entre modelos; cada una describe una propiedad distinta del mismo sistema. La decisión clínica requiere conocer umbral, población y consecuencias del error.

Du y colaboradores (19) realizaron un metaanálisis de algoritmos de aprendizaje profundo aplicados a laringoscopia que incluyó nueve estudios y más de 106.000 imágenes endoscópicas. La sensibilidad agrupada para el diagnóstico de cáncer laríngeo fue de 95%, mientras que la especificidad fue de 96% y el área bajo la curva alcanzó aproximadamente 0,99. Estos resultados evidencian el potencial diagnóstico de la tecnología, aunque los autores también reconocen la necesidad de continuar investigando su utilidad clínica.

De manera complementaria, Marrero-Gonzalez y colaboradores (20) analizaron aplicaciones de IA en endoscopia, análisis de voz e histopatología. Para la endoscopia asistida por IA, la sensibilidad combinada para clasificar lesiones benignas y malignas fue de aproximadamente 91%, mientras que la sensibilidad para detección de lesiones también fue cercana al 91%.

Una revisión más reciente de Alabdalhussein y colaboradores (21), que incluyó 15 estudios y 17.559 pacientes, encontró una sensibilidad agrupada de 78% y especificidad de 86% para la detección de cáncer laríngeo. Esta diferencia respecto de otros metaanálisis demuestra que el rendimiento de los modelos depende considerablemente del diseño de los estudios, población analizada, definición del desenlace, tipo de imágenes y estrategia algorítmica.

Por tanto, la evidencia disponible no debería interpretarse como la existencia de un único rendimiento diagnóstico universal de la IA. Más apropiadamente, los estudios demuestran que determinados sistemas pueden alcanzar un rendimiento elevado bajo condiciones específicas. (17)

Uno de los desafíos más importantes en la detección temprana consiste en identificar lesiones que todavía no constituyen un carcinoma invasivo. La leucoplasia de las cuerdas vocales representa un ejemplo especialmente relevante debido a que puede corresponder a cambios benignos, displasia o carcinoma. (22)

Tie y colaboradores (22) desarrollaron un modelo de aprendizaje profundo basado en multi-instance learning para distinguir leucoplasias benignas y malignas utilizando imágenes de luz blanca y NBI. El estudio incluyó miles de imágenes procedentes de múltiples instituciones y realizó validación externa. El área bajo la curva alcanzó 0,868 en la validación interna y 0,884 en la externa. Además, la evaluación con laringólogos mostró que la asistencia de IA podía mejorar determinadas medidas diagnósticas y la consistencia de los especialistas.

Este tipo de investigación resulta especialmente relevante porque representa un cambio desde la simple clasificación “cáncer/no cáncer” hacia una herramienta destinada a resolver un problema clínico concreto: determinar qué lesión aparentemente sospechosa requiere mayor atención diagnóstica. (22)

La tabla 2 organiza metaanálisis seleccionados: poblaciones y estimaciones, con indicación del alcance de cada componente y de las fuentes que sustentan su interpretación.

Tabla 2

Metaanálisis seleccionados: poblaciones y estimaciones

Síntesis	Unidad declarada	Resultado informado
Du (19)	9 estudios; 106.175 imágenes	Sensibilidad 95%; especificidad 96%
Alabdulhussein (21)	15 estudios; 17.559 pacientes	Sensibilidad 78%; especificidad 86%
Marrero-Gonzalez (20)	Modalidades diferenciadas	Sensibilidad endoscópica 91%
Żurek (17)	Endoscopia laríngea	Variación entre modelos y tareas

Fuente: No se suman cohortes ni se calcula una nueva estimación agrupada.

La IA podría, por tanto, actuar como un sistema de triaje. Ante una imagen endoscópica, el algoritmo podría señalar zonas sospechosas y asignar una probabilidad de malignidad, permitiendo al especialista decidir si se requiere biopsia, seguimiento estrecho u otra evaluación. (22)

Otra aplicación relevante es la segmentación automática de lesiones. En lugar de limitarse a indicar si existe o no una lesión, un algoritmo de segmentación identifica los píxeles correspondientes al área patológica. (6)

Sampieri y colaboradores (6) desarrollaron un sistema de aprendizaje profundo denominado SegMENT-Plus destinado a delimitar en tiempo real los bordes del cáncer laríngeo en imágenes de luz blanca y NBI. El modelo presentó resultados comparables a los de residentes de otorrinolaringología y fue probado en conjuntos de datos externos, mostrando una capacidad de generalización prometedora.

La segmentación podría adquirir importancia durante la planificación quirúrgica. Una delimitación más precisa de los márgenes tumorales podría facilitar la evaluación de la extensión superficial de una lesión y apoyar la planificación de procedimientos endoscópicos conservadores. (6)

Los metaanálisis muestran estimaciones diferentes y no necesariamente contradictorias.

Du y colaboradores incluyeron nueve estudios y 106.175 imágenes, con sensibilidad y especificidad agrupadas de 95% y 96%; Alabdalussein y colaboradores integraron 15 estudios y 17.559 pacientes, con 78% y 86%. Los denominadores, tareas y criterios de selección impiden atribuir la diferencia únicamente al año o a la calidad de los algoritmos. (19,21)

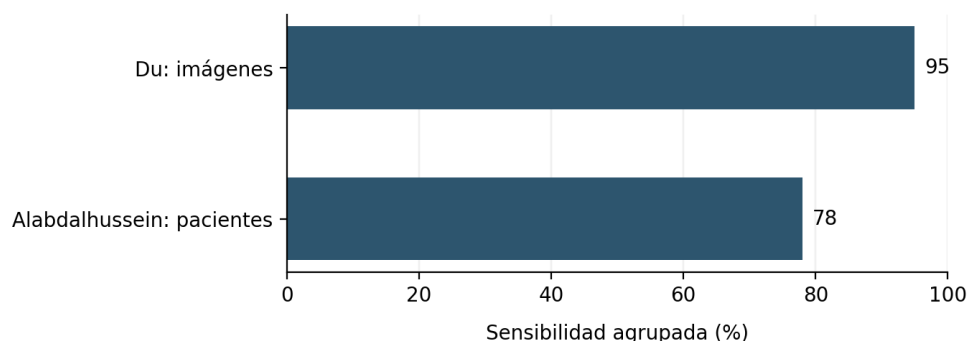
La revisión de Marrero-Gonzalez y colaboradores distinguió endoscopia, voz e histopatología y estimó sensibilidad de 91% para clasificación endoscópica de lesiones benignas y malignas. La síntesis de Żurek y colaboradores también documentó resultados altos, con variación entre estudios. El número de revisiones no representa replicaciones independientes si varias incluyen los mismos modelos o cohortes. (20,17)

La revisión de alcance sobre videómica amplía el mapa de tareas y requisitos técnicos. Su utilidad es identificar qué se ha estudiado y cómo, no proporcionar una recomendación universal de adopción. La transición desde clasificación retrospectiva hacia análisis de vídeo exige resolver estabilidad, latencia y comportamiento ante imágenes no interpretables. (23)

El gráfico 2 presenta estimaciones publicadas de sensibilidad en dos metaanálisis. Se conserva la unidad y el contexto de los datos para distinguir la descripción publicada de las implicaciones propuestas.

Gráfico 2

Estimaciones publicadas de sensibilidad en dos metaanálisis



Fuente: elaboración propia. (19,21). Corpus y definiciones diferentes; no es comparación directa entre algoritmos.

Las diferencias entre síntesis no acreditan una mejora o deterioro histórico de la tecnología. Cambian criterios, tareas, unidades y estudios incluidos; además, pueden existir cohortes compartidas. El gráfico muestra variación de estimaciones publicadas para impedir una cifra universal de rendimiento, sin proponer una combinación nueva.

Sin embargo, todavía es necesario demostrar que una mejor segmentación computacional se traduce efectivamente en mejores resultados oncológicos. Un algoritmo puede delimitar una lesión con elevada precisión matemática sin necesariamente modificar las tasas de resección completa, recurrencia o preservación funcional. (6)

Una de las aplicaciones más atractivas de la IA es su incorporación directamente al flujo de la videolaringoscopia. En este escenario, el algoritmo analiza cada fotograma y proporciona una señal visual al especialista cuando identifica una región potencialmente sospechosa. (9)

El procesamiento en tiempo real podría tener varias ventajas. Primero, permitiría identificar lesiones durante el procedimiento sin necesidad de revisar posteriormente miles de imágenes. Segundo, podría ayudar a disminuir el riesgo de pasar por alto lesiones pequeñas. Tercero, podría servir como apoyo educativo para especialistas en formación. (9)

No obstante, la velocidad computacional no debe confundirse con utilidad clínica. Un sistema de tiempo real debe minimizar falsos positivos, evitar alertas excesivas y proporcionar resultados suficientemente estables durante los movimientos de la cámara, cambios de iluminación, secreciones, sangre o desenfoque. (9)

El rendimiento de la IA puede mejorar cuando el algoritmo tiene acceso a diferentes modalidades de imagen. La luz blanca aporta información sobre la arquitectura y coloración de la mucosa, mientras que NBI puede proporcionar información adicional sobre patrones vasculares. (22)

La combinación de ambas modalidades puede ser particularmente útil para lesiones con características ambiguas. Los estudios de Tie y colaboradores (22) representan un ejemplo de esta estrategia multimodal.

La futura evolución de los sistemas podría integrar simultáneamente imágenes endoscópicas, video, información clínica, antecedentes de exposición a tabaco y alcohol, edad, síntomas y resultados histopatológicos. Sin embargo, esta integración debe realizarse cuidadosamente para evitar modelos que aprendan asociaciones espurias. (22)

La tabla 3 organiza riesgos de sesgo y controles de evaluación, con indicación del alcance de cada componente y de las fuentes que sustentan su interpretación.

Tabla 3

Riesgos de sesgo y controles de evaluación

Riesgo	Consecuencia	Control
Fotogramas del mismo paciente	Fuga entre entrenamiento y prueba	Separación por paciente
Imágenes seleccionadas	Rendimiento poco transportable	Casos consecutivos y exclusiones
Verificación diferencial	Etiquetas diagnósticas sesgadas	Referencia y seguimiento claros
Umbral post hoc	Optimismo en resultados	Umbral predefinido y validación

Fuente: Elaboración propia basada en (12,13,24,25,10,11).

Un algoritmo podría, por ejemplo, asociar una determinada cámara, hospital o grupo poblacional con el diagnóstico en lugar de aprender verdaderamente las características de la lesión. Por ello, la validación externa y multicéntrica resulta fundamental. (22)

A pesar del crecimiento de la evidencia, existen limitaciones importantes. La primera es la heterogeneidad de las bases de datos. Los estudios emplean diferentes cámaras, resoluciones, sistemas de iluminación, protocolos de adquisición y criterios de clasificación. (24)

La segunda es la dependencia de estudios retrospectivos. Los algoritmos suelen entrenarse utilizando imágenes seleccionadas después de conocer el diagnóstico. En la práctica clínica, el sistema debe funcionar antes de conocer el resultado histopatológico y en una población donde la prevalencia de cáncer sea diferente. (24)

La tercera limitación es el riesgo de sobreajuste. Cuando imágenes del mismo paciente aparecen simultáneamente en los conjuntos de entrenamiento y validación, el rendimiento puede parecer artificialmente elevado. La división correcta debe realizarse idealmente a nivel del paciente y, cuando sea posible, mediante validación externa en instituciones independientes. (24)

El estudio de Tie y colaboradores que combinó luz blanca y NBI mediante aprendizaje de múltiples instancias reportó áreas bajo la curva de 0,868 en validación interna, 0,884 en externa y 0,825 en vídeos prospectivos. El conjunto de vídeo incluyó 50 exploraciones y su intervalo fue amplio; la variación ilustra la importancia de evaluar el sistema en condiciones cercanas al uso previsto. (22)

La publicación de Wang y colaboradores sobre aprendizaje de múltiples instancias constituye un trabajo relacionado, pero no debe confundirse con el estudio anterior por compartir autores y tema. La trazabilidad requiere asociar cada resultado al artículo y

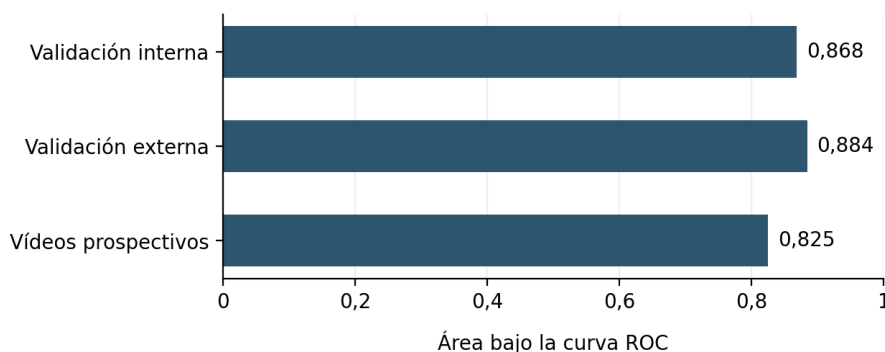
conjunto correspondiente; duplicar cohortes o intercambiar sus métricas puede producir una impresión artificial de validación independiente. (26)

La leucoplasia es una apariencia clínica con posibles sustratos diferentes. Un modelo puede ayudar a priorizar evaluación, pero la probabilidad estimada depende de la composición de lesiones con la que se desarrolló. Las investigaciones de clasificación y de riesgo de transformación deben distinguir un diagnóstico presente de un evento futuro, que necesita seguimiento temporal apropiado. (16,27)

El gráfico 3 presenta discriminación del modelo multimodal de leucoplasia. Se conserva la unidad y el contexto de los datos para distinguir la descripción publicada de las implicaciones propuestas.

Gráfico 3

Discriminación del modelo multimodal de leucoplasia



Fuente: elaboración propia. (22). Conjuntos diferentes; vídeo: 50 exploraciones. No demuestra beneficio clínico.

El desempeño en vídeos fue menor en estimación puntual, pero los conjuntos difieren y su comparación requiere considerar precisión y composición. La validación externa elevada no garantiza idéntico comportamiento en tiempo real. Para implementación interesa estabilidad de alertas y decisiones del especialista, además de discriminación en imágenes seleccionadas.

La cuarta limitación es la generalización. Un modelo entrenado principalmente con pacientes de una determinada región geográfica puede no mantener el mismo desempeño en otras poblaciones. (24)

Finalmente, existe el problema de la interpretabilidad. Los especialistas necesitan conocer por qué el sistema considera sospechosa una imagen. Las técnicas de visualización de regiones de interés pueden ayudar, pero todavía no sustituyen una explicación clínica completa. (24)

La utilización de IA para diagnóstico médico exige altos estándares de seguridad. Una clasificación incorrecta puede producir consecuencias clínicas relevantes: un falso negativo podría retrasar el diagnóstico de un cáncer, mientras que un falso positivo podría generar biopsias, ansiedad y procedimientos innecesarios. (14)

Por ello, los estudios deben reportar no solamente exactitud, sino también sensibilidad, especificidad, valores predictivos, curvas ROC, intervalos de confianza y características de la población estudiada. (14)

Las recomendaciones STARD-AI publicadas en 2025 destacan precisamente la necesidad de describir de forma transparente las bases de datos, el funcionamiento del algoritmo, los métodos de evaluación y potenciales sesgos. (14)

Asimismo, las extensiones CONSORT-AI y SPIRIT-AI proporcionan estándares para evaluar intervenciones clínicas que incorporan inteligencia artificial, haciendo énfasis en la interacción entre profesionales y sistemas de IA y en el reporte de errores del algoritmo (28,29).

La IA debe considerarse, por tanto, una herramienta de apoyo y no una sustitución automática del otorrinolaringólogo. La decisión diagnóstica definitiva debe integrar hallazgos endoscópicos, antecedentes, exploración clínica, imagen complementaria cuando corresponda y confirmación histopatológica. (14)

La tabla 4 organiza implementación clínica supervisada propuesta, con indicación del alcance de cada componente y de las fuentes que sustentan su interpretación.

Tabla 4

Implementación clínica supervisada propuesta

Etapas	Evidencia requerida	Decisión
Validación local silenciosa	Desempeño por población y equipo	Ajustar o detener
Piloto asistido	Usabilidad, fallos y decisiones	Mantener supervisión
Comparación clínica	Diagnóstico, biopsias y tiempos	Valorar beneficio y daño
Seguimiento	Versión, cambios y degradación	Revisar o retirar

Fuente: Propuesta basada en (14,28,29,5); no representa despliegue ya evaluado.

La evolución futura de esta tecnología probablemente estará marcada por modelos multimodales y sistemas capaces de analizar no solamente fotografías, sino secuencias completas de video. El concepto de videomics plantea la posibilidad de extraer información temporal y espacial de grandes cantidades de imágenes endoscópicas. (5)

La segmentación automática también podría adquirir mayor relevancia, especialmente en el contexto de cirugía endoscópica. Además, los sistemas podrían evolucionar hacia plataformas capaces de identificar lesiones, delimitar sus bordes, estimar el nivel de sospecha y comparar los hallazgos con estudios previos. (5)

La investigación publicada en 2026 muestra que el campo está avanzando hacia modelos más clínicamente orientados, incluyendo sistemas destinados simultáneamente a la detección de cáncer y a la estimación del riesgo de transformación maligna de lesiones laríngeas (16).

Sin embargo, la prioridad no debería ser únicamente desarrollar algoritmos con mayor exactitud. El objetivo principal debe ser demostrar que su utilización mejora resultados clínicos relevantes: detección en estadios tempranos, reducción de biopsias innecesarias, menor tiempo hasta el diagnóstico, mejor selección de pacientes para procedimientos y, finalmente, mejores resultados funcionales y oncológicos. (5)

SegMENT-Plus se desarrolló con 3.933 imágenes de 557 pacientes y se examinó en dos conjuntos externos de 156 y 200 imágenes. La mediana del coeficiente Dice fue 0,83 y la velocidad de inferencia, 25,6 fotogramas por segundo. Tales resultados apoyan factibilidad técnica, pero no establecen márgenes histológicos negativos ni reducción de recurrencia; la extensión superficial visible es solo una parte del problema quirúrgico.

(6)

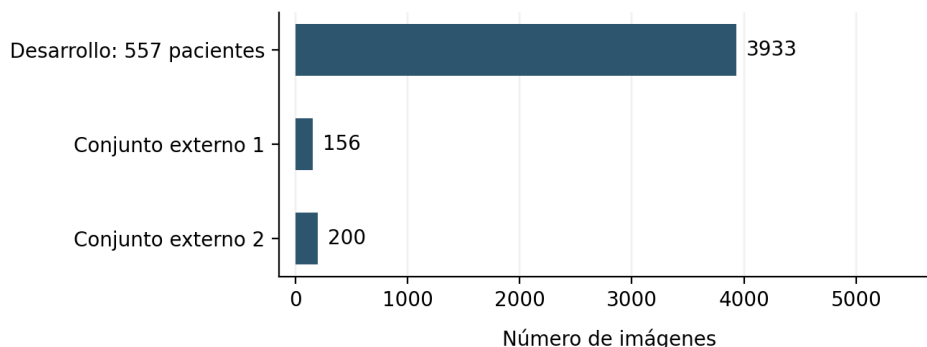
Wellenstein y colaboradores mostraron posibilidad de procesamiento rápido durante endoscopia. La velocidad debe evaluarse junto con latencia del sistema completo, persistencia de alertas y calidad de imagen; un número de fotogramas por segundo medido en condiciones experimentales no garantiza estabilidad con todos los equipos clínicos. (9)

La revisión de impacto clínico en endoscopia del tracto aerodigestivo superior y el estudio prospectivo de Hu y colaboradores acercan la pregunta al escenario asistencial. La clasificación como prospectivo mejora la pertinencia temporal, pero no elimina sesgos de selección ni demuestra por sí sola que la asistencia cambie desenlaces importantes. La propuesta exige vincular desempeño técnico con decisiones y consecuencias. (5,30)

El gráfico 4 presenta imágenes utilizadas en desarrollo y validación de SegMENT-Plus. Se conserva la unidad y el contexto de los datos para distinguir la descripción publicada de las implicaciones propuestas.

Gráfico 4

Imágenes utilizadas en desarrollo y validación de SegMENT-Plus



Fuente: elaboración propia. (6). Imágenes no equivalen a pacientes independientes; el desarrollo comprende varias etapas.

El conjunto de desarrollo contiene muchas más imágenes que cada validación externa. Esa asimetría no invalida el estudio, pero obliga a examinar incertidumbre y diversidad de pacientes; múltiples imágenes de una misma persona no equivalen a observaciones independientes. La factibilidad de segmentación debe complementarse con evaluación clínica prospectiva.

Discusión

La evidencia acumulada permite considerar la inteligencia artificial como una tecnología con potencial significativo para complementar la detección endoscópica del cáncer de laringe. Los modelos de aprendizaje profundo han demostrado capacidad para reconocer patrones asociados con lesiones malignas y premalignas, y los metaanálisis recientes presentan resultados diagnósticos elevados. (5)

Sin embargo, la magnitud de los resultados varía entre estudios. Las diferencias entre las sensibilidades agrupadas reportadas por Du y colaboradores (19) y Alabdalhussein y colaboradores (21) evidencian que el rendimiento depende de las características de los datos, la definición de la tarea diagnóstica y el diseño de validación.

Un aspecto particularmente importante es que el objetivo clínico no debe limitarse a detectar carcinomas evidentes. La verdadera oportunidad de la IA está en identificar

lesiones tempranas y potencialmente precursoras, cuando la intervención puede contribuir a preservar la función laríngea y reducir la progresión de la enfermedad. (5)

La integración con NBI representa una vía especialmente prometedora. La IA puede aprovechar características vasculares que requieren experiencia para su interpretación, aumentando potencialmente la uniformidad diagnóstica. Del mismo modo, los sistemas de segmentación pueden proporcionar información adicional sobre extensión superficial. (5)

No obstante, la IA no elimina la necesidad de biopsia e histopatología. La confirmación microscópica continúa siendo necesaria para establecer el diagnóstico definitivo de malignidad. La función más realista de la IA en el corto plazo es actuar como herramienta de asistencia, priorización y segunda lectura. (5)

Otro elemento esencial es la colaboración multidisciplinaria. El desarrollo de estos sistemas requiere la participación conjunta de otorrinolaringólogos, oncólogos, patólogos, ingenieros biomédicos, científicos de datos y especialistas en metodología clínica. (5)

Los estudios futuros deberían utilizar cohortes multicéntricas, protocolos estandarizados de adquisición de imágenes, validación externa y diseños prospectivos. También deben analizar el desempeño del sistema cuando se utiliza conjuntamente con especialistas de diferentes niveles de experiencia. (5)

La principal conclusión crítica es que exactitud y utilidad clínica pertenecen a niveles distintos. Un modelo puede reconocer patrones y, aun así, no mejorar el diagnóstico si señala lesiones ya evidentes, produce alertas que se ignoran o no cambia el tiempo hasta la biopsia. La evaluación debe describir qué decisión espera modificar y comparar atención habitual con atención asistida bajo condiciones realistas. (5,14)

La selección de imágenes constituye un riesgo metodológico central. Si se excluyen fotogramas borrosos, zonas difíciles y exploraciones negativas complejas, el conjunto de prueba se aleja de la consulta habitual. Un algoritmo destinado a apoyar al especialista necesita reconocer incertidumbre o abstenerse ante datos insuficientes; presentar una categoría siempre, aunque falte información, puede aumentar confianza injustificada.

La división de datos debe realizarse por paciente y, cuando proceda, por centro o período. Fotogramas del mismo vídeo comparten rasgos anatómicos, iluminación y lesión; distribuirlos entre entrenamiento y prueba permite reconocer similitudes sin demostrar generalización. Los estándares de reporte sobre imágenes y modelos ayudan a identificar esa dependencia, aunque el cumplimiento formal no garantiza ausencia de sesgo. (13,12)

La referencia diagnóstica también puede ser imperfecta. La histología debe corresponder a la lesión y momento examinados; cuando solo las lesiones sospechosas reciben biopsia, aparece verificación diferencial. Clasificar el resto como benigno sin seguimiento suficiente puede distorsionar sensibilidad y especificidad; por ello, debe describirse qué ocurrió con todos los pacientes y no únicamente con las imágenes disponibles para entrenamiento. (10,11)

La interpretación de las métricas requiere considerar consecuencias. Un falso negativo puede retrasar evaluación de una lesión relevante y un falso positivo puede aumentar procedimientos y ansiedad. La elección del umbral debe responder al uso previsto y a capacidad asistencial, mientras que el área bajo la curva resume discriminación a través de umbrales sin indicar cuál es clínicamente aceptable.

Los valores predictivos cambian con prevalencia y selección de pacientes. Un sistema probado en una serie enriquecida con cáncer no puede trasladar directamente su

probabilidad posterior a una consulta de disfonía con otra composición. La calibración debe examinar si probabilidades estimadas corresponden a frecuencias observadas; una buena discriminación no asegura estimaciones individuales bien calibradas. (12,24,25)

La validación externa es necesaria, pero su significado depende de la independencia real. Centros que comparten equipos, protocolos o población pueden constituir una prueba limitada de transporte; la validación temporal y multicéntrica debería incluir condiciones relevantes para el destino. La falta de significación al comparar conjuntos pequeños no acredita equivalencia de rendimiento ni ausencia de degradación.

La evaluación humano-IA requiere diseñar comparaciones justas. Deben considerarse experiencia, orden de lectura, memoria de imágenes y tiempo disponible; un mismo caso repetido con y sin asistencia puede introducir aprendizaje. También interesa estudiar qué ocurre cuando el modelo se equivoca y si especialistas mantienen capacidad de discrepancia, porque la aceptación automática puede neutralizar beneficios de una buena herramienta.

Las diferencias entre luz blanca y NBI no se resuelven necesariamente añadiendo modalidades. La combinación puede aportar información, pero también aumentar requisitos de adquisición y restringir aplicabilidad a centros equipados. El desarrollo debe describir qué sucede si falta una modalidad; comparar un modelo multimodal con una lectura limitada sin justificar condiciones puede favorecer artificialmente al sistema. (22,26)

La segmentación exige una lectura independiente de la clasificación. Una máscara que coincide con anotaciones expertas representa extensión visible, no prueba de profundidad, invasión o margen quirúrgico seguro. Para evaluar beneficio quirúrgico se requieren desenlaces relacionados con resección, función y recurrencia, además de explicar variabilidad entre anotadores y correspondencia con histopatología. (6)

La implementación propuesta comienza con evaluación silenciosa del modelo en datos locales y continúa con asistencia supervisada limitada. El paso entre fases debería depender de desempeño, seguridad y usabilidad definidos previamente; no se presupone que un resultado retrospectivo autorice despliegue general. Los marcos DECIDE-AI, SPIRIT-AI y CONSORT-AI permiten ordenar evaluación clínica y transparencia de ensayos posteriores. (14,29,28)

La gobernanza debe identificar versión, responsables, cambios y mecanismos de retirada. Un ajuste del modelo o del equipo endoscópico puede alterar desempeño; por ello, el seguimiento necesita denominadores, registro de fallos y revisión periódica. El consentimiento y la protección de vídeos requieren especial cuidado porque contienen información clínica e imágenes potencialmente identificables, aun cuando no aparezca el rostro.

La equidad se relaciona con representatividad clínica y acceso. Diferencias de comorbilidad, lesiones tratadas previamente, anatomía o calidad de adquisición pueden afectar resultados; la evaluación debe examinar grupos relevantes y no limitarse a una cifra global. En centros con menos recursos, una herramienta no debería aumentar dependencia de derivaciones inexistentes ni sustituir inversión en diagnóstico definitivo. Las limitaciones de esta revisión incluyen selección narrativa y dependencia de reportes heterogéneos. No se recalcularon métricas a partir de bases individuales ni se aplicó una evaluación completa y duplicada de riesgo de sesgo. El análisis utiliza los estándares como criterios críticos y conserva la diferencia entre resultados publicados y propuesta de implementación; no demuestra eficacia de un producto específico.

La investigación futura debe priorizar estudios prospectivos con pacientes consecutivos, comparación del proceso asistencial y evaluación de daños. Son desenlaces relevantes el tiempo hasta diagnóstico confirmado, lesiones omitidas, biopsias innecesarias y

resultados funcionales, junto con costes y carga de trabajo. La supervivencia requiere seguimiento y diseños adecuados; no puede inferirse de un área bajo la curva elevada.

Conclusiones

La inteligencia artificial aplicada al análisis de imágenes endoscópicas constituye una de las líneas emergentes con mayor potencial dentro de la innovación diagnóstica en laringología. Las redes neuronales convolucionales y otros modelos de aprendizaje profundo han demostrado capacidad para identificar lesiones laríngeas, diferenciar lesiones benignas y malignas, detectar lesiones precursoras y delimitar bordes tumorales. (17)

La evidencia disponible hasta 2026 muestra resultados diagnósticos elevados en diferentes escenarios, aunque no existe todavía suficiente homogeneidad metodológica para asumir que estos valores serán reproducibles de manera universal en la práctica clínica. (17)

La combinación de endoscopia de luz blanca, NBI y algoritmos de IA podría mejorar la detección de alteraciones sutiles y contribuir a estandarizar la interpretación de las imágenes. Además, el procesamiento en tiempo real ofrece posibilidades para integrar sistemas de asistencia directamente en la videolaringoscopia. (17)

No obstante, la implementación clínica debe sustentarse en estudios prospectivos, multicéntricos y externamente validados. La transparencia metodológica, la evaluación de sesgos, la interpretabilidad, la seguridad y la interacción humano-máquina deben formar parte del proceso de validación. (17)

En consecuencia, la IA debe entenderse actualmente como una herramienta complementaria al juicio clínico especializado. Su mayor contribución potencial no reside en reemplazar al otorrinolaringólogo, sino en aumentar su capacidad para detectar

tempranamente lesiones sospechosas, priorizar pacientes y mejorar la precisión y consistencia de la evaluación endoscópica. (17)

La IA endoscópica laríngea muestra capacidad para apoyar clasificación y delimitación de lesiones, con variación importante entre tareas y conjuntos. Su contribución a detección temprana clínicamente útil depende de validación independiente, interacción segura con especialistas y seguimiento de decisiones y desenlaces. La matriz propuesta establece estos requisitos sin atribuir al rendimiento retrospectivo beneficios oncológicos todavía no demostrados.

Referencias bibliográficas

1. Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. *Nature reviews. Disease primers*. 2020;6(1):92. <https://doi.org/10.1038/s41572-020-00224-3>
2. Igissin N, Zatonskikh V, Telmanova Z, Tulebaev R, Moore M. Laryngeal Cancer: Epidemiology, Etiology, and Prevention: A Narrative Review. *Iranian journal of public health*. 2023;52(11):2248-2259. <https://doi.org/10.18502/ijph.v52i11.14025>
3. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. <https://doi.org/10.1038/nature14539>
4. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Medical image analysis*. 2017;42:60-88. <https://doi.org/10.1016/j.media.2017.07.005>
5. Wilmes CMLH, BSc AG, Marres HAM, Wellenstein DJ, van den Broek GB. A Systematic Review of the Clinical Impact of Implementing Artificial Intelligence in Upper Aerodigestive Tract Endoscopy. *Head & neck*. 2025;47(11):2998-3018. <https://doi.org/10.1002/hed.28213>

6. Sampieri C, Azam MA, Ioppi A, Baldini C, Moccia S, Kim D, et al. Real-Time Laryngeal Cancer Boundaries Delineation on White Light and Narrow-Band Imaging Laryngoscopy with Deep Learning. *The Laryngoscope*. 2024;134(6):2826-2834. <https://doi.org/10.1002/lary.31255>
7. Xiong H, Lin P, Yu JG, Ye J, Xiao L, Tao Y, et al. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine*. 2019;48:92-99. <https://doi.org/10.1016/j.ebiom.2019.08.075>
8. Yao P, Usman M, Chen YH, German A, Andreadis K, Mages K, et al. Applications of Artificial Intelligence to Office Laryngoscopy: A Scoping Review. *The Laryngoscope*. 2022;132(10):1993-2016. <https://doi.org/10.1002/lary.29886>
9. Wellenstein DJ, Woodburn J, Marres HAM, van den Broek GB. Detection of laryngeal carcinoma during endoscopy using artificial intelligence. *Head & neck*. 2023;45(9):2217-2226. <https://doi.org/10.1002/hed.27441>
10. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ (Clinical research ed.)*. 2015;351:h5527. <https://doi.org/10.1136/bmj.h5527>
11. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature medicine*. 2025;31(10):3283-3289. <https://doi.org/10.1038/s41591-025-03953-8>
12. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ (Clinical research ed.)*. 2024;385:e078378. <https://doi.org/10.1136/bmj-2023-078378>

13. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiology. Artificial intelligence*. 2024;6(4):e240300. <https://doi.org/10.1148/ryai.240300>
14. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ (Clinical research ed.)*. 2022;377:e070904. <https://doi.org/10.1136/bmj-2022-070904>
15. Ni XG, He S, Xu ZG, Gao L, Lu N, Yuan Z, et al. Endoscopic diagnosis of laryngeal cancer and precancerous lesions by narrow band imaging. *The Journal of laryngology and otology*. 2011;125(3):288-296. <https://doi.org/10.1017/s0022215110002033>
16. Kang Y, Hu B, Xu K, Cai L, Li W, Sun S, et al. A Clinically Inspired Deep Learning Framework for Laryngeal Cancer Detection and Malignancy Transformation Risk Assessment. *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery*. 2026;174(1):173-184. <https://doi.org/10.1002/ohn.70075>
17. Żurek M, Jasak K, Niemczyk K, Rzepakowska A. Artificial Intelligence in Laryngeal Endoscopy: Systematic Review and Meta-Analysis. *Journal of clinical medicine*. 2022;11(10):2752. <https://doi.org/10.3390/jcm11102752>
18. Lam SWS, Lee MH, Dorosan M, Altonji S, Tan HK, Lee WT. Use of a Preliminary Artificial Intelligence-Based Laryngeal Cancer Screening Framework for Low-Resource Settings: Development and Validation Study. *JMIR formative research*. 2025;9:e66110. <https://doi.org/10.2196/66110>
19. Du S, Guo J, Huang D, Liu Y, Zhang X, Lu S. Diagnostic accuracy of deep learning-based algorithms in laryngoscopy: a systematic review and meta-analysis.

- European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery. 2025;282(1):351-360. <https://doi.org/10.1007/s00405-024-09049-2>
20. Marrero-Gonzalez AR, Diemer TJ, Nguyen SA, Camilon TJM, Meenan K, O'Rourke A. Application of artificial intelligence in laryngeal lesions: a systematic review and meta-analysis. European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery. 2025;282(3):1543-1555. <https://doi.org/10.1007/s00405-024-09075-0>
21. Alabdalhussein A, Al-Khafaji MH, Al-Busairi R, Al-Dabbagh S, Khan W, Anwar F, et al. Artificial Intelligence in Laryngeal Cancer Detection: A Systematic Review and Meta-Analysis. Current oncology (Toronto, Ont.). 2025;32(6):338. <https://doi.org/10.3390/curroncol32060338>
22. Tie CW, Li DY, Zhu JQ, Wang ML, Wang JH, Chen BH, et al. Multi-Instance Learning for Vocal Fold Leukoplakia Diagnosis Using White Light and Narrow-Band Imaging: A Multicenter Study. The Laryngoscope. 2024;134(10):4321-4328. <https://doi.org/10.1002/lary.31537>
23. Ioppi A, Bellini E, Salvetta MS, Marchi F, di Maria D, Peretti G, et al. Videomics and artificial intelligence in endoscopic diagnosis of laryngeal lesions: mapping current evidence through a scoping review. Acta otorhinolaryngologica Italica : organo ufficiale della Societa italiana di otorinolaringologia e chirurgia cervico-facciale. 2026;46(Suppl. 1):S19-S33. <https://doi.org/10.14639/0392-100x-suppl.1-46-2026-a1967>

24. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of internal medicine*. 2019;170(1):51-58. <https://doi.org/10.7326/m18-1376>
25. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Annals of internal medicine*. 2019;170(1):W1-W33. <https://doi.org/10.7326/m18-1377>
26. Wang ML, Tie CW, Wang JH, Zhu JQ, Chen BH, Li Y, et al. Multi-instance learning based artificial intelligence model to assist vocal fold leukoplakia diagnosis: A multicentre diagnostic study. *American journal of otolaryngology*. 2024;45(4):104342. <https://doi.org/10.1016/j.amjoto.2024.104342>
27. Xiong M, Luo JW, Ren J, Hu JJ, Lan L, Zhang Y, et al. Applying Deep Learning with Convolutional Neural Networks to Laryngoscopic Imaging for Automated Segmentation and Classification of Vocal Cord Leukoplakia. *Ear, nose, & throat journal*. 2024;1455613241275341. <https://doi.org/10.1177/01455613241275341>
28. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature medicine*. 2020;26(9):1364-1374. <https://doi.org/10.1038/s41591-020-1034-x>
29. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature medicine*. 2020;26(9):1351-1363. <https://doi.org/10.1038/s41591-020-1037-7>

30. Hu R, Liu X, Zhang Y, Arthur C, Qin D. Comparison of clinical nasal endoscopy, optical biopsy, and artificial intelligence in early diagnosis and treatment planning in laryngeal cancer: a prospective observational study. *Frontiers in oncology*. 2025;15:1582011. <https://doi.org/10.3389/fonc.2025.1582011>