Predicción de la Supervivencia en Pacientes con Cáncer de Estómago: Integración de Características Clínicas, Genéticas y Análisis de Imágenes para el Apoyo en la Toma de Decisiones Clínicas

William Andrés López León

Karem Dayana Meneses Ramírez

Eliana Liseth Parra Barrera

Códigos 9014227, 9013913, 9012723

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director: Fabian Tobar Tosse, Ph.D

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, 08 DE DICIEMBRE 2025

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

El cáncer gástrico es una enfermedad neoplásica con una etiología multifactorial, caracterizada por el crecimiento descontrolado de células tumorales en la mucosa del estómago, y es responsable de aproximadamente 768,000 muertes a nivel mundial. Afecta principalmente a hombres mayores de 70 años en Asia, Europa y América Latina, y aunque ha habido una disminución de casos en América del Norte y Europa Occidental, sigue siendo el quinto cáncer más común a nivel mundial [1], [2]. La enfermedad es a menudo asintomática y se diagnostica en etapas avanzadas, lo que limita las posibilidades de cura. La supervivencia de los pacientes varía según factores como el momento del diagnóstico, características demográficas y clínicas, y tratamientos recibidos. En tal contexto, la Ciencia de Datos se presenta como una herramienta poderosa para mejorar el diagnóstico y la medicina de precisión en el cáncer gástrico [1], [3].

De manera puntual, la Ciencia de Datos, mediante el empleo de algoritmos de aprendizaje automático y la capacidad de procesar grandes volúmenes de datos heterogéneos, permite integrar información procedente de múltiples fuentes, desde características demográficas y clínicas hasta datos genómicos e imágenes histopatológicas digitalizadas, lo que contribuye a una comprensión más integral de la biología y evolución del tumor, así como de la respuesta del paciente al tratamiento. La naturaleza multifactorial del cáncer gástrico y la variabilidad en la respuesta al tratamiento hacen que los enfoques predictivos basados en datos sean indispensables para superar las limitaciones actuales en el pronóstico y el diseño de terapias personalizadas.

La detección temprana del cáncer gástrico es esencial para mejorar los resultados clínicos, dado que la tasa de supervivencia a cinco años suele ser inferior al 40% [4], [5]. Países con sistemas de salud avanzados, supera el 18% de supervivencia a cinco años [6].

Con este sustento, la presente investigación tiene por objetivo de este proyecto integrar las características demográficas, clínicas, genéticas y el análisis de imágenes histológicas para desarrollar un modelo predictivo que apoye la toma de decisiones clínicas en el manejo de pacientes con cáncer de estómago y la predicción de la sobrevida con un enfoque multidimensional para apoyar la precisión pronóstica, el diagnóstico, la atención y el tratamiento de los pacientes.

2. CONTEXTUALIZACIÓN DEL PROYECTO

2.1 Definición del problema

2.2 Planteamiento del problema

El cáncer de estómago es una enfermedad en la que células cancerosas se forman en el revestimiento del estómago [7], [8]. A nivel mundial, 768, 000 muertes fueron atribuidas a cáncer de estómago [9]. Esta enfermedad afecta principalmente a hombres mayores de 70 años en Asia, Europa y América Latina [10]. Aunque en los últimos 60 años se ha observado una disminución en el número de casos en América del Norte y Europa Occidental, el cáncer de estómago sigue siendo el quinto cáncer más común a nivel mundial, con proyección a aumentar los casos a futuro [10], [11].

Esta enfermedad a menudo es asintomática y suele diagnosticarse en etapas avanzadas cuando las posibilidades de cura son mínimas. Los indicios más comunes que presenta incluyen dificultad para tragar, debilidad, indigestión, vómitos, pérdida de peso, sensación de saciedad temprana y anemia por deficiencia de hierro [12]. Sin embargo, estos síntomas no son específicos, lo que hace que el diagnóstico temprano sea poco común y contribuye a que el 80% de los casos se detecten tardíamente. El análisis definitivo requiere procedimientos endoscópicos, toma de muestras y estudios histopatológicos de la biopsia de la lesión, lo que puede (en algunos casos) retrasar el inicio del tratamiento [13], [14]. Debido a ello la supervivencia del paciente con cáncer de estómago es variada según las características de cada caso, el momento de diagnósticos, los factores demográficos y clínicos. En países como Estados Unidos y Japón han descrito tasas de supervivencia de 5 años desde el 42.9% al 71,1% para el cáncer de estómago, respectivamente [1], [15].

La supervivencia en cáncer de estómago depende, entre otras cosas, de características del paciente, factores tumorales, y tratamientos recibidos, así como la identificación y manejo multifactorial adecuado pueden ayudar a mejorar los resultados clínicos de los pacientes. De igual manera, la edad, el sexo y la etnia han sido asociadas a una mayor sobrevida, tal es el caso de pacientes más jóvenes, mujeres y ciertas etnias (asiáticos) han presentado mejores tasas de sobrevida [16], [17]. Han de tenerse en cuenta también otras características asociadas al estado nutricional y ausencia de comorbilidades [18], [19]. El estadio del tumor y la diferenciación histológica al momento del diagnóstico son factores críticos, los pacientes con la detección localizada, tienen mayor supervivencia que aquellos con una distribución metastásica. Al igual, la diferenciación de la patología ha sido asociada con una mejor supervivencia comparada con tumores pobremente diferenciados [16], [20].

Estudios han determinado que el grado de diferenciación tumoral, la metástasis ganglionar regional y el estadio patológico posoperatorio eran factores de riesgo independientes para la supervivencia general a 5 años en pacientes con cáncer de estómago avanzado [20], [21].

La profundidad de la invasión del tumor primario y la rentabilidad ganglionar se asociaron significativamente con la supervivencia general en los pacientes con cáncer de estómago [21]. La invasión serosal (ypT4) se asoció con una alta tasa de recurrencia peritoneal, y se deben considerar los ensayos de terapia intraperitoneal dirigidos a estos pacientes [21].

Comprender los factores pronósticos del cáncer de estómago avanzado antes de comenzar la quimioterapia es importante para determinar estrategias de tratamiento personalizadas. Sin embargo, los detalles de la quimioterapia y el pronóstico de los pacientes con cáncer de estómago avanzado han cambiado con el tiempo y el entorno [22]. Un estudio multicéntrico en Japón, con una cohorte de 1025 pacientes con tratamiento de quimioterapia sistémica para cáncer de estómago avanzado en 12 instituciones de salud, determinó que la mediana de supervivencia global y supervivencia libre de progresión de la quimioterapia de primera línea fue de 11,8 meses (IC 95%10,8–12,3 meses) y 6,3 meses (IC 95%, 5,9–6,9 meses), respectivamente. Asociando a factores como edad < 40 años, estado funcional ≥2, sin gastrectomía, tipo histológico difuso, albúmina <3,6, fosfatasa alcalina ≥300, creatinina ≥1,0 y cociente neutrófilos/linfocitos > 3,0, y el tratamiento trastuzumab mostraron una mejor supervivencia que los pacientes que no lo utilizaron (16,1 meses frente a 11,1 meses; p = 0,0005) [22].

La ciencia de datos ha tenido un impacto profundo en casi todos los aspectos del cáncer de estómago, desde la mejora del diagnóstico hasta la medicina de precisión [23], [24], [25], [26]. Estudios recientes han mostrado la que procedimiento tradicionales complejos como la inspección visual de láminas histopatológicas de ganglios linfáticos, para calcular el número de ganglios linfáticos metastásicos (MLNs), han sido analizadas con el fin de identificar ganglios linfáticos y regiones tumorales, y luego descubrir la proporción de área tumoral a área de MLNs, y que después del entrenamiento, el rendimiento de detección de tumores los modelos son comparable al de patólogos experimentados con logró en rendimiento similar en dos cohortes de validación independientes de cáncer de estómago. Indicando que los modelos de aprendizaje profundo podrían ayudar no solo a los patólogos en la detección de ganglios linfáticos con metástasis, sino también a los oncólogos en la exploración de nuevos factores pronósticos, especialmente aquellos que son difíciles de calcular manualmente [27]. Por consiguiente, el aprendizaje automático en el análisis de imágenes para el cáncer gástrico ofrece herramientas para mejorar la predicción de la supervivencia y la identificación de factores pronósticos, aunque su implementación clínica aún enfrenta desafíos significativos, los modelos de aprendizaje automatizado para predecir la supervivencia y pueden identificar factores influyentes en la predicción [24], [25], [26].

Actualmente, existen grandes cantidades de información clínica, genómica e imágenes histopatológicas disponibles en reconocidos repositorios científicos de datos de cáncer que posibilitan tener cada vez características más detalladas sobre los casos de pacientes diagnosticados con cáncer de estómago (Cancer Genome Atlas (TCGA), Gene Expression

Omnibus (GEO), European Genome-phenome Archive (EGA), ArrayExpress, NCBI, cBioPortal for Cancer Genomics, Open Targets Platform, entre otras). Lo cual, presenta la oportunidad de desarrollar un enfoque multidimensional de análisis de datos en el que se integren diferentes fuentes de información que permita potencializar la precisión pronostica de la supervivencia de los pacientes, mejorar el diagnóstico, indicar mejoras en la atención, el tratamiento, la calidad de vida y evitar la recurrencia de la enfermedad. Por consiguiente, los modelos de aprendizaje automatizado para predecir la supervivencia pueden identificar factores influyentes en la predicción. En este sentido, el problema se plantea en términos de determinar modelos de predicción por aprendizaje automático para determinar los factores asociados a la sobrevida en un enfoque mixto con el análisis de imágenes diagnosticas.

2.3 Formulación del problema

¿Cómo predecir la supervivencia de pacientes con cáncer de estómago integrando características clínicas, genéticas y analítica de imágenes histológicas mediante la aplicación de la ciencia de datos?

2.4 Objetivos

2.4.1 Objetivo General

Predecir la supervivencia de pacientes diagnosticados con cáncer de estómago, integrando características clínicas, genéticas y analítica de imágenes histológicas para apoyo en la toma de decisiones clínicas referentes al manejo de pacientes.

2.4.2 Objetivos Específicos

1. Preparar los datos clínicos, genéticos e histopatológicos de los pacientes para entrenar un modelo de predicción.

2. Integrar las características clínicas, genéticas e imágenes histopatológicas en una base de datos unificada que permita predecir la supervivencia en pacientes con cáncer de estómago.

3. Desarrollar modelos que permitan el predecir de supervivencia de pacientes con cáncer de estómago de forma precisa y fiable.

4. Evaluar la calidad del modelo mediante métricas de rendimiento, asegurando su efectividad y aplicabilidad en el contexto clínico.

3. MARCO DE REFERENCIA

3.1. Marco teórico

3.1.1. Fundamentos del cáncer de estomago

3.1.1.1. Definición y subtipos histológicos

El cáncer gástrico es un problema de salud pública de gran importancia. El cáncer gástrico es la quinta neoplasia más frecuente y una de las principales causas de mortalidad por cáncer a nivel mundial [28], [29]. Se ubica en el tercer puesto entre las causas de muerte por cáncer reportadas con 783.000 muertes cada año. La incidencia del cáncer gástrico varía considerablemente por regiones geográficas. Las tasas más elevadas de cáncer gástrico se registran en Asia Oriental (Japón, Corea del Sur, China), Europa del Este y América Latina (especialmente en Colombia, Chile y Perú), mientras que las tasas más bajas se observan en América del Norte y África Subsahariana. En Japón y Corea del Sur, el cáncer gástrico constituye la neoplasia maligna más común entre los hombres y es una de las principales causas de mortalidad por cáncer. En China, continúa siendo una causa significativa de muerte por cáncer. En Estados Unidos y Europa Occidental, la incidencia ha disminuido progresivamente durante las últimas décadas, aunque se ha identificado un incremento relativo en los casos de aparición temprana (<50 años) [29].

El cáncer de estómago es una neoplasia maligna originada en el epitelio del estómago, caracterizada por una marcada heterogeneidad clínica, histológica y molecular. Es una de las neoplasias que se encuentran dentro de las más frecuentes a nivel mundial y representa una importante carga de morbimortalidad a las personas que lo padecen, representando el cuarto lugar en frecuencia y el segundo lugar en mortalidad por cáncer en todo el mundo. Esta neoplasia que se forma en los tejidos que revisten el estómago. La mayoría comienza en las células de la mucosa, denominados adenocarcinomas y representan aproximadamente el 90% de los casos de cáncer de estómago [30]. La progresión del cáncer gástrico incluye varias etapas que se muestran en la Tabla I [30], [31].

Tabla I. Progresión histopatológica hacia cáncer de estómago

Etapa	Descripción breve	Características clave	Riesgo de progresión
Estómago normal	Mucosa gástrica sana y funcional	Glándulas íntegras; producción adecuada de ácido y enzimas	Base fisiológica (sin riesgo aumentado)
Gastritis crónica no atrófica (GCN-A)	Inflamación crónica, frecuentemente por *H. pylori*	Mucosa inflamada; glándulas aún conservadas	Riesgo bajo–moderado si persiste la infección
Gastritis crónica atrófica	Pérdida progresiva de glándulas (atrofia)	Disminuye acidez; adelgazamiento mucoso; disbiosis	Riesgo moderado; favorece etapas siguientes

Metaplasia intestinal (MI)	Sustitución por epitelio tipo intestinal (completo o incompleto)	Marcador de alto riesgo; cambios adaptativos persistentes	Riesgo alto (↑ si metaplasia incompleta)
Displasia (neoplasia intraepitelial)	Atipia celular y desorganización sin invasión	Bajo o alto grado; precursora inmediata de cáncer	Riesgo muy alto (especialmente alto grado)
Cáncer gástrico temprano	Invasión a lámina propia y/o submucosa sin alcanzar muscular	Alta curabilidad con resección endoscópica/cirugía	Riesgo de progresión si no se trata; buen pronóstico si tratado
Cáncer gástrico avanzado	Invasión muscular, ganglionar y/o metástasis a distancia	Pronóstico desfavorable; requiere terapias multimodales	Máximo riesgo; alta mortalidad si tardío

La clasificación de histopatológica establece una distinción fundamental que conecta el origen y el comportamiento tumoral: el tipo intestinal, que se desarrolla a partir de un proceso secuencial de atrofia y metaplasia en la mucosa gástrica [32], [33], formando estructuras glandulares y siendo más común en ancianos; y el tipo difuso, que se caracteriza por células mal cohesionadas que infiltran la pared gástrica sin formar glándulas, con una fuerte asociación a factores genéticos y un peor pronóstico. Si bien estos adenocarcinomas dominan el panorama, existen otros tumores menos comunes que también se originan en el estómago, como los linfomas, los tumores del estroma gastrointestinal (GIST) y los tumores neuroendocrinos [33], [34], [35].

3.1.1.2. Epidemiología global y regional y factores de riesgo

Los factores de riesgo más relevantes para el cáncer gástrico incluyen la infección crónica por Helicobacter pylori, antecedentes familiares, consumo de tabaco, dietas ricas en sal y pobres en frutas y verduras, obesidad y consumo excesivo de alcohol. También influyen ciertos hábitos alimentarios como la ingesta frecuente de alimentos ahumados o curados, y condiciones ambientales como la falta de refrigeración adecuada o el acceso limitado a agua potable. Entre los factores genéticos destacan el grupo sanguíneo A, la anemia perniciosa y síndromes hereditarios como el cáncer colorrectal no polipoide y Li-Fraumeni. La infección por H. pylori, en particular, se asocia con gastritis crónica, úlceras pépticas, linfoma MALT y adenocarcinoma gástrico. La prevención primaria se orienta a reducir factores causales como el tabaquismo y los hábitos alimentarios de riesgo, mientras que la prevención secundaria se centra en la detección temprana y el seguimiento de individuos con alto riesgo [36], [37].

Alimentación: Las dietas saladas incrementan el riesgo de patologías gástricas. El alto consumo de sal altera la mucosa gástrica, favorece la colonización y virulencia del *H. pylori*, aumenta la exposición a compuestos carcinogénicos y estimula respuestas inflamatorias que llevan a mutaciones en el epitelio gástrico. La ingesta de frutas y verduras está directamente relacionada con el desarrollo de ciertas enfermedades. Una ingesta

inadecuada o deficiente puede actuar como un factor predisponente, mientras que una ingesta alta se asocia con un riesgo reducido. Adicionalmente, el uso de técnicas específicas al cocinar alimentos, como hornear, freír o asar carnes, entre otros, provoca la formación de compuestos N-nitrosos, que están vinculados a un mayor riesgo de cáncer gástrico [38].

Estilos de vida: El tabaquismo influye de forma importante en la aparición de distintos tipos de cáncer, lo hace también en el cáncer gástrico. El consumo de alcohol aumenta la ingesta de nitrosaminas al causar una respuesta inflamatoria crónica por los metabolitos y citocinas del etanol. El riesgo es mayor en personas que consumen 4 o más bebidas al día, aumentando un 24% en quienes consumen más de 50 g de alcohol diarios, en comparación con los no bebedores. La obesidad está vinculada al cáncer gástrico. La grasa abdominal y la producción de metabolitos activos como el factor de crecimiento similar a la insulina aumentan el riesgo de cáncer en la unión esofagogástrica y cardias. En contraste, la actividad física reduce el riesgo de desarrollar esta enfermedad [37], [38].

Genética: Los polimorfismos relacionados con el cáncer gástrico incluyen los genes de citocinas IL1RNVNTR, CYP19A1CYPE1, NAT2 M1 y XRCC1 194. También se han identificado muchos loci asociados a la aparición del cáncer gástrico. Los síndromes hereditarios causan del 1 al 3% de los cánceres gástricos. Entre ellos están el cáncer gástrico difuso hereditario, una neoplasia autosómica dominante causada por una alteración molecular que afecta las adherencias intercelulares debido a la falta de E-cadherina. Otros síndromes relacionados son la poliposis adenomatosa familiar y el síndrome de Peutz-Jeghers [29], [38].

Historial familiar de cáncer: se considera un factor de riesgo al tener un familiar de primer grado con cáncer gástrico. Esto puede ser consecuencia del entorno compartido, como en casos donde se transmite la infección por *H. Pylori* de padres a hijos, o de aspectos genéticos previamente mencionados.

Antecedentes médicos: esta categoría incluye los aspectos relacionados con las condiciones médicas del paciente, como patologías infecciosas y el uso de medicamentos, que pueden actuar como factores de riesgo. El grupo sanguíneo A+ se asocia con una mayor presencia de cáncer gástrico, mientras que en personas con grupo A- es menos frecuente. Además, antecedentes patológicos como úlceras gástricas, gastritis atrófica y metaplasia intestinal son factores de riesgo, así como cirugías estomacales debido a la disminución del ácido gástrico post-intervención. También hay relación entre el cáncer gástrico y factores reproductivos, como la menopausia y la edad del primer parto o menarca. La menopausia aumenta el riesgo de esta neoplasia [39].

Factores sociodemográficos: El riesgo de cáncer gástrico aumenta con la edad. El 1% de los casos ocurrieron en personas entre 20 y 34 años y el 20% entre 75 y 84 años. Los hombres tienen mayor riesgo que las mujeres, posiblemente por factores ambientales y fisiológicos, como la edad fértil y la menopausia que pueden influir en este riesgo debido a las hormonas femeninas. La raza blanca es la que mayor probabilidad de enfermedad. El nivel

socioeconómico bajo representa un factor de riesgo importante, pero por su relación con otros factores como el nivel de salud, la alimentación y elementos ambientales, como la infección por *H. pylori* [39].

3.1.1.3. Diagnóstico, estadificación y pronóstico clínico

El cáncer gástrico suele manifestarse con síntomas inespecíficos, siendo los síntomas más frecuentemente observados la epigastralgia, la plenitud abdominal, la dispepsia, la llenura precoz y menos frecuentes son las náuseas y los vómitos. Posteriormente, en etapas más avanzadas se puede presentar pérdida del apetito, síntomas constitucionales, sangrados y pérdida de peso [40].

En su diagnóstico inicial, se utiliza la endoscopia digestiva, aunque otras modalidades como la ecografía endoscópica, las radiografías convencionales, la tomografía computarizada, la resonancia magnética y la endoscopia virtual se emplean también. La endoscopia digestiva alta incluye toma de biopsias para confirmación histológica, y la estadificación requiere estudios de imagen (TC, ultrasonido endoscópico, PET, laparoscopia). El diagnóstico del cáncer gástrico se realiza principalmente mediante endoscopia superior y radiografía por contraste. La endoscopia gastrointestinal superior es preferida para diagnosticar el cáncer gástrico, ya que permite visualizar directamente la mucosa gástrica y obtener biopsias para identificar lesiones precancerosas como atrofia gástrica, metaplasia intestinal o displasia gástrica.

El examen histopatológico establece el tipo y grado del tumor. Para determinar la extensión de la enfermedad, se emplea la clasificación TNM del AJCC (American Joint Committee on Cancer por sus siglas en ingles), que evalúa la profundidad de invasión del tumor (T), el compromiso ganglionar (N) y la presencia de metástasis a distancia (M). La estadificación clínica incluye, además de la endoscopía, estudios de imagen como tomografía computarizada (TC), PET/CT, ultrasonido endoscópico (EUS) y ocasionalmente laparoscopía diagnóstica, sobre todo en tumores localmente avanzados.

El cáncer gástrico es diagnosticado en etapas avanzadas de la enfermedad lo cual es crítico porque posee una alta capacidad metastásica (estructuras intraabdominales y el hígado). Numerosos sistemas de clasificación histológica han sido propuestos para el cáncer gástrico, que incluye la clasificación de la Organización Mundial de la Salud y de Bormann, sin embargo, el sistema de clasificación de Lauren es ampliamente aceptado, el cual consiste en un sistema histopatológico para categorizar el adenocarcinoma gástrico en dos tipos principales, el intestinal y el difuso. El primero, se determina según el sitio proximal/cardia o distal/no cardia y su subtipo histológico principal intestinal formador de glándulas, asociado a factores ambientales como infección por *Helicobacter pylori*, dieta alta en sal y bajo consumo de frutas y verduras y el tipo difuso que es por células poco cohesionadas, vinculado a predisposición genética y de comportamiento más agresivo [9]. La histología del cáncer gástrico es heterogénea, pero el adenocarcinoma, en sus variantes intestinal y

difusa, constituye la gran mayoría de los casos, con una gama de subtipos menos frecuentes que requieren reconocimiento especializado para un manejo óptimo [40].

La estadificación del cáncer gástrico es un proceso crucial que determina la extensión y la diseminación de la enfermedad en el momento del diagnóstico. Es fundamental para guiar el tratamiento (decisiones sobre cirugía, quimioterapia, radioterapia o terapias dirigidas), establecer el pronóstico (la etapa es uno de los predictores más importantes de la supervivencia del paciente), la comunicación (lenguaje estandarizado para que los profesionales de la salud), y la investigación (comparar los resultados de los tratamientos y estudios clínicos de manera consistente). La estadificación completa a menudo requiere una combinación de estudios de imagen (endoscopia con biopsia, tomografía computarizada, PET-CT, laparoscopia diagnóstica) y evaluación patológica de la pieza quirúrgica y los ganglios linfáticos resecados. La estadificación más utilizada es el sistema TNM, desarrollado por el American Joint Committee on Cancer (AJCC) y la Union for International Cancer Control (UICC). Este sistema evalúa tres componentes principales:

- **T (Tumor):** Describe el tamaño del tumor primario y hasta dónde se ha extendido en las capas de la pared del estómago.
- **N (Nódulo/Ganglio):** Indica si el cáncer se ha diseminado a los ganglios linfáticos cercanos al estómago y, de ser así, cuántos ganglios están afectados.
- **M (Metástasis):** Determina si el cáncer se ha diseminado a partes distantes del cuerpo (metástasis a distancia), como el hígado, los pulmones, los huesos o los ganglios linfáticos alejados.

Una vez que se determinan las categorías T, N y M, se combinan para asignar una etapa general al cáncer, que va desde la etapa 0 hasta la etapa IV como se presenta en la Tabla II.

Tabla II. Componentes y etapas del cáncer de estómago

COMPONENTE	CLASIFICACIÓN	DESCRIPCIÓN
T (TUMOR PRIMARIO)	Tis	Carcinoma in situ (células cancerosas solo en la capa más superficial del revestimiento del estómago, sin invadir la lámina propia).
	T1	El tumor invade la lámina propia, la muscularis mucosae o la submucosa.
	T1a	El tumor invade la lámina propia o la muscularis mucosae.
	T1b	El tumor invade la submucosa.
	T2	El tumor invade la muscularis propria (capa muscular principal).
	T3	El tumor invade la subserosa (capa más externa del estómago antes del peritoneo), sin invadir el peritoneo visceral.

	T4	El tumor invade la serosa (peritoneo visceral) o estructuras adyacentes.
	T4a	El tumor invade la serosa.
	T4b	El tumor invade estructuras adyacentes (bazo, hígado, diafragma, etc.).
N (GANGLIOS LINFÁTICOS REGIONALES)	N0	No hay metástasis en los ganglios linfáticos regionales.
	N1	Metástasis en 1 a 2 ganglios linfáticos regionales.
	N2	Metástasis en 3 a 6 ganglios linfáticos regionales.
	N3	Metástasis en 7 o más ganglios linfáticos regionales.
	N3a	Metástasis en 7 a 15 ganglios linfáticos regionales.
	N3b	Metástasis en 16 o más ganglios linfáticos regionales.
M (METÁSTASIS A DISTANCIA)	M0	No hay metástasis a distancia.
	M1	Hay metástasis a distancia (incluyendo ganglios linfáticos no regionales, siembra peritoneal, metástasis en órganos distantes como hígado, pulmón, etc.).

Fuente: Elaboración propia siguiendo las referencias [41], [42], [43]

La combinación de las categorías T, N y M define las etapas clínicas del cáncer gástrico:

Etapa 0 (Tis, N0, M0): Carcinoma in situ. Las células cancerosas están solo en la capa más interna del revestimiento del estómago y no se han diseminado. Es altamente curable; etapa I (T1, N0, M0): El tumor es pequeño y solo ha crecido en las capas más internas del estómago, sin afectar ganglios ni tener metástasis a distancia; etapa II (Varias combinaciones de T, N, M0): El tumor ha crecido más profundamente en la pared del estómago, o ha afectado algunos ganglios linfáticos cercanos, pero aún no hay metástasis a distancia; etapa III (Varias combinaciones de T, N, M0): El cáncer se ha extendido más profundamente en la pared del estómago y/o ha afectado más ganglios linfáticos cercanos, pero todavía no hay metástasis a distancia; etapa IV (Cualquier T, cualquier N, M1): El cáncer se ha diseminado a otras partes del cuerpo, lejos del estómago. Esta es la etapa más avanzada y generalmente incurable.

3.1.1.4. Desafíos actuales en el manejo del cáncer de estómago

El cáncer gástrico requiere una aproximación multidisciplinaria, integración de nuevas tecnologías diagnósticas, personalización terapéutica basada en biomarcadores y una mejor inclusión de poblaciones vulnerables, todo ello en el contexto de una investigación clínica coordinada y global. Los desafíos actuales en el manejo del cáncer gástrico incluyen la heterogeneidad tumoral, la estadificación insuficiente, la optimización de estrategias

quirúrgicas y sistémicas, la atención a poblaciones vulnerables, y la necesidad de avanzar en medicina de precisión y acceso equitativo a cuidados multidisciplinarios como se presenta en la Tabla III.

Tabla III. Desafíos actuales en el manejo del cáncer gástrico

Desafío	Descripción	Implicaciones
Heterogeneidad biológica y molecular	El cáncer gástrico presenta una gran variabilidad en histología, localización, perfil genético y respuesta a tratamientos.	Dificulta la estratificación de pacientes, la selección de tratamientos óptimos y la implementación de estrategias personalizadas, llevando a resultados inconsistentes en ensayos clínicos.
Limitaciones en la estadificación y diagnóstico	Las técnicas radiológicas convencionales no detectan con precisión la enfermedad metastásica oculta, especialmente en el peritoneo.	Causa una subestadificación, lo que lleva a decisiones terapéuticas subóptimas. La laparoscopia diagnóstica y la citología peritoneal mejoran la precisión, pero no son de uso universal.
Desafíos quirúrgicos y multimodales	Existen controversias sobre la extensión óptima de la linfadenectomía (D2), el uso de técnicas mínimamente invasivas y la función de la quimiorradioterapia adyuvante.	Se necesita adaptar las técnicas quirúrgicas y los esquemas de tratamiento, especialmente con el aumento de tumores de localización proximal y subtipo difuso.
Resistencia a terapias sistémicas	La resistencia primaria y adquirida limita la eficacia de terapias dirigidas (anti-HER2, antiangiogénicos) e inmunoterapia.	La variabilidad en la expresión de biomarcadores como HER2 y PD-L1 dificulta la selección de pacientes y la durabilidad de la respuesta.
Envejecimiento poblacional y comorbilidades	Una porción significativa de pacientes son adultos mayores con riesgo de toxicidad, comorbilidades y fragilidad.	La falta de ensayos clínicos específicos y la poca integración de la valoración geriátrica dificultan la toma de decisiones individualizadas y la optimización del tratamiento.
Acceso desigual a innovaciones	La disponibilidad de diagnósticos avanzados, terapias innovadoras y equipos multidisciplinarios varía considerablemente entre regiones.	Afecta la equidad en el manejo y los resultados del tratamiento.
Necesidad de biomarcadores y medicina de precisión	Se requiere el desarrollo e integración de biomarcadores predictivos y plataformas multi-ómicas.	El avance hacia la medicina personalizada está en fases iniciales de validación clínica.
Complejidad del microambiente tumoral	El microambiente tumoral es complejo, compuesto por células inmunitarias, vasos sanguíneos y estroma, lo que afecta la respuesta a tratamientos.	Dificulta la predicción de la respuesta a la inmunoterapia y otras terapias sistémicas, y es un área de investigación activa.

Desafíos en la implementación de la medicina de precisión	La aplicación de análisis genéticos avanzados presenta desafíos logísticos y económicos.	Los altos costos, el tiempo de los análisis y la variabilidad en la interpretación de los biomarcadores dificultan su implementación rutinaria.
Gestión de la toxicidad y calidad de vida	Los tratamientos (quimioterapia, terapias dirigidas, inmunoterapia) tienen efectos secundarios significativos.	El manejo integral de la toxicidad y la preservación de la calidad de vida de los pacientes son desafíos constantes y cruciales.
Desafíos en detección temprana	Los síntomas del cáncer gástrico en estadios iniciales son a menudo inespecíficos o inexistentes.	Esto conduce a diagnósticos tardíos. La implementación de programas de cribado y la concienciación pública son fundamentales, pero complejas.

Fuente propia usando estas referencias [44], [45], [46], [47], [48], [49], [50], [51].

3.1.2. Fundamentos de supervivencia en el cáncer de estómago

3.1.2.1. Importancia del análisis de supervivencia en el cáncer de estómago

El análisis de supervivencia se utiliza para determinar tasas de supervivencia global y específica por cáncer, así como para evaluar el impacto de variables clínicas, patológicas y terapéuticas sobre el pronóstico. Este método estadístico no solo permite estimar la probabilidad de supervivencia de los pacientes a lo largo del tiempo, sino que también es crucial para evaluar el pronóstico, personalizar los tratamientos y diseñar programas de seguimiento eficaces [52], [53], [54]. A través de técnicas como el análisis de Kaplan-Meier o la supervivencia condicional, se pueden identificar factores pronósticos clave como el estadio del tumor, la edad, el estado nutricional y ciertos biomarcadores que influyen en el resultado.

Estos datos son esenciales para tomar decisiones clínicas informadas y adaptar las estrategias terapéuticas a las necesidades de cada paciente. Además, el análisis de supervivencia ayuda a medir la eficacia de las intervenciones y a monitorear la mejora en las tasas de supervivencia, las cuales han aumentado significativamente en los últimos años, especialmente en los estadios tempranos de la enfermedad [55]. Esta herramienta es indispensable para mejorar los resultados clínicos y la calidad de vida de quienes enfrentan esta patología. El uso de modelos multivariados como el de Cox ha permitido identificar factores independientes de mal pronóstico, como la diferenciación tumoral, la afectación ganglionar y el estadio patológico posquirúrgico, lo que orienta la selección de tratamientos y el seguimiento personalizado [56].

3.1.2.2. Definición de supervivencia y tiempo hasta el evento

La supervivencia se define como el tiempo transcurrido desde un punto de referencia como el diagnóstico o el inicio del tratamiento hasta que ocurre un evento de interés, usualmente la muerte, aunque también puede corresponder a recurrencia, progresión tumoral u otros desenlaces clínicos relevantes. Cuando el evento no se observa dentro del periodo de

seguimiento, el individuo se clasifica como censurado, una característica fundamental de los estudios de supervivencia [57]. La censura puede ser: *por la derecha* (el evento no ocurre antes del fin del estudio), *por la izquierda* (el evento ocurre antes de la inclusión) o *por intervalo* (solo se conoce que ocurrió dentro de un rango temporal).

El análisis de supervivencia constituye un conjunto de métodos estadísticos que se utilizan en diversos campos, donde se analiza el tiempo transcurrido hasta que ocurre un evento de interés, no solo si ocurre, permitiendo evaluar factores que influyen en el riesgo asociado mediante una función de supervivencia (Ecuación 1) y una función de riesgo (Ecuación 2) [58], [59], [60].

$$S(t) = P(T > t)$$

Ecuación 1. Formula de la estimación del riesgo en función de la sobrevida.

$$h(t) = \frac{P(t \leq T < t + dt | T \geq t)}{(T - t)} = \frac{f(t)}{S(t)}$$

Ecuación 2. Formula de la estimación del riesgo en función del riesgo.

$h(t)$*: Función de riesgo. Representa la tasa instantánea a la que ocurre un evento (por ejemplo, la muerte o la falla de un equipo) en el tiempo t, dado que el evento no ha ocurrido antes de ese momento. No es una probabilidad, sino una tasa.*

$P(t \leq T < t + dt | T \geq t)$*: Probabilidad condicional. Esta es la probabilidad de que el evento ocurra en un intervalo de tiempo muy pequeño [t,t+dt), dado que el sujeto u objeto ha sobrevivido (es decir, el evento no ha ocurrido) hasta el tiempo t.*

T: Variable de tiempo. Es una variable aleatoria que representa el tiempo hasta que ocurre el evento de interés.

t: Tiempo. Es un punto específico en el tiempo.

dt: Intervalo de tiempo infinitesimal.

f(t): Función de densidad de probabilidad. Describe la probabilidad relativa de que el evento ocurra exactamente en el tiempo t.

S(t): Función de supervivencia. Representa la probabilidad de que un sujeto u objeto sobreviva más allá del tiempo t, es decir, la probabilidad de que el evento no haya ocurrido en o antes del tiempo t. Se define como $S(t) = P(T \geq t)$.

La función de supervivencia expresa la probabilidad de no experimentar el evento antes del tiempo t, mientras que la función de riesgo describe la tasa instantánea a la cual ocurre el

evento en t, dado que no ha ocurrido previamente. Ambas funciones permiten estimar pronósticos y comparar grupos terapéuticos o clínicos en oncología [58], [59], [60].

En cáncer gástrico, los principales desenlaces analizados son la supervivencia global (OS) y la supervivencia libre de progresión (PFS), indicadores esenciales para evaluar tratamientos y estrategias de manejo [61], [62]. La mejora progresiva de las técnicas quirúrgicas, la adopción de esquemas perioperatorios modernos y la selección individualizada para tratamientos multimodales han modificado sustancialmente los patrones de supervivencia en los últimos años [63], [64]. En cohortes contemporáneas, la supervivencia relativa a 5 años puede alcanzar hasta 71.4%, con diferencias marcadas según estadio tumoral: hasta 89.7% en estadios I–III, frente a ~29% en estadio IV [65]. La supervivencia media en pacientes tratados con quimioterapia o quimiorradioterapia perioperatoria se sitúa entre 46 y 49 meses, con tasas de supervivencia a 5 años del 44–46% [66].

Diversos estudios identifican factores pronósticos independientes asociados a peor evolución, entre ellos la diferenciación histológica, la presencia de metástasis ganglionares, el estadio patológico postoperatorio y la miosteatosis. La localización tumoral, la profundidad de invasión y el tamaño tumoral también influyen en la supervivencia, aunque no siempre se mantienen como predictores independientes en modelos multivariados. Estos hallazgos se resumen en la Tabla IV, donde se destacan los principales determinantes del pronóstico y los avances recientes en resultados de supervivencia [67], [68].

Tabla IV. Supervivencia y factores pronósticos en el cáncer gástrico

Aspecto de la Supervivencia	Hallazgo Clave y Tendencias	Relevancia e Implicaciones
Supervivencia Global	Mejora progresiva en los últimos años. Tasas de supervivencia relativa a 5 años de hasta el 71.4% en centros de alto volumen.	Atribuido a la optimización de técnicas quirúrgicas, esquemas perioperatorios modernos y tratamientos multimodales.
Supervivencia por Estadio	La supervivencia a 5 años varía significativamente: hasta 89.7% en estadios I-III vs. ~29% en estadio IV.	El estadio tumoral sigue siendo el principal determinante del pronóstico.
Eficacia de Tratamientos Sistémicos	Mediana de supervivencia global de 46-49 meses con quimioterapia/quimiorradioterapia perioperatoria. Tasas de supervivencia a 5 años de 44-46%.	No se observan diferencias significativas en supervivencia entre los principales esquemas, lo que subraya la importancia de la selección individualizada.
Factores Pronósticos Independientes	Peor supervivencia asociada a: grado de diferenciación, metástasis ganglionares, estadio patológico postoperatorio y miosteatosis.	Factores como la miosteatosis emergen como predictores significativos de mortalidad, especialmente en pacientes quirúrgicos.

Estrategias de Seguimiento	La vigilancia intensiva es crucial en los primeros 7-8 años post-cirugía (estadios II y III) debido al mayor riesgo de recaída.	El seguimiento regular con endoscopia y tomografía se asocia a una reducción de la mortalidad a largo plazo y a una mejor supervivencia post-recaída.
Impacto de la Calidad de la Atención	La centralización de la cirugía y el acceso a tratamientos multimodales benefician a la mayoría de los pacientes.	El impacto es más notable en estadios no metastásicos, resaltando la importancia de la centralización y la atención especializada.

3.1.2.3. Características clínicas y genéticas relevantes en el análisis de supervivencia

La supervivencia oncológica depende de integrar características clínicas y genéticas para una estratificación pronóstica más precisa. Factores como estadio tumoral, respuesta al tratamiento, metástasis, histología y alteraciones genéticas (mutaciones, LOH, SNPs funcionales, carga mutacional) influyen directamente en el pronóstico y en la selección terapéutica [67].

En cáncer gástrico, el estadio patológico (pTNM) es el principal determinante de supervivencia. La invasión tumoral (T), el compromiso ganglionar (N) y la presencia de metástasis (M) se asocian de forma independiente con la supervivencia global y específica [67], [69]. Además, la invasión tumoral profunda (T3 o mayor), incluso sin metástasis ganglionares, y la diferenciación pobre del tumor son indicadores de mal pronóstico. La presencia de invasión linfovascular o perineural también se asocia con mayor riesgo de recurrencia y menor supervivencia, particularmente en estadios avanzados [70], [71].

En estadio I, se asocian a peor supervivencia la edad avanzada (≥65 años), pT2 y diámetro tumoral ≥5 cm. En supervivientes tras gastrectomía, edad ≥80 años y NLR ≥2.7 reducen la supervivencia global; estadio III y MCV elevado se relacionan con menor supervivencia específica y no específica por cáncer [72].

Entre factores clínico-quirúrgicos, el tipo de cirugía, la resección de órganos adyacentes y, sobre todo, la radicalidad (márgenes negativos) determinan supervivencia prolongada [73]. El sexo masculino y la localización proximal se han vinculado a peor pronóstico en subgrupos. En estadios avanzados, metástasis ganglionares y el estadio patológico postoperatorio son predictores robustos de mal pronóstico . La respuesta a neoadyuvancia y la ausencia de complicaciones mayores influyen, aunque menos que los factores anatomopatológicos clásicos.

En conjunto, estos hallazgos sustentan los enfoques de modelado predictivo y se resumen en la Tabla V [73].

Tabla V. Factores pronósticos clínicos, patológicos y genéticos de supervivencia en el cáncer gástrico

Tipo de Factor	Factor Pronóstico	Descripción y Relevancia para la Supervivencia
I. Clínicos y Patológicos	Estadio Patológico (pTNM)	El estadio es el principal determinante pronóstico. La profundidad de invasión tumoral (T), la afectación ganglionar (N) y la metástasis a distancia (M) se asocian de manera independiente con la supervivencia. La invasión tumoral profunda (pT3 o mayor) es un factor de mal pronóstico incluso en pacientes sin afectación ganglionar (N0) [69].
	Diferenciación Histológica	Los tumores poco diferenciados presentan un peor pronóstico.
	Invasión Linfovascular y Perineural	Se asocia con un mayor riesgo de recurrencia y menor supervivencia, especialmente en estadios intermedios y avanzados.
	Calidad y Extensión de la Cirugía	La disección insuficiente de ganglios linfáticos (<16) se asocia con peor supervivencia. La radicalidad de la resección (márgenes negativos) es fundamental para la supervivencia a largo plazo.
	Respuesta a Terapia Neoadyuvante	La respuesta patológica completa o parcial a la quimioterapia/quimiorradioterapia preoperatoria influye en la supervivencia.
	Factores Demográficos y del Paciente	Edad avanzada (≥65 años en estadio I, o ≥80 años en supervivientes a largo plazo), sexo masculino y localización tumoral proximal se asocian con un peor pronóstico en algunos subgrupos.
	Biomarcadores Hematológicos	Un índice neutrófilo/linfocito (NLR) elevado (≥2,7) y un volumen corpuscular medio (MCV) alto se asocian con menor supervivencia en supervivientes a largo plazo.
	Tamaño Tumoral	Un diámetro tumoral ≥5 cm en estadio I es un factor independiente de peor supervivencia.
	Complicaciones Postoperatorias	La ausencia de complicaciones postoperatorias mayores puede influir positivamente en la supervivencia.
II. Genéticos y Moleculares	Variantes Intrónicas	Variantes específicas en genes como MGMT (rs12268840) y STARD3/PGAP3 (rs9972882) son predictores independientes. Se asocian con un mayor riesgo de recaída y metástasis, influyendo en la progresión tumoral.

MiARN en la Carcinogénesis Gástrica

Los microARN (miARN) intervienen en la carcinogénesis gástrica como reguladores post-transcripcionales de genes relacionados con la proliferación, apoptosis, invasión y resistencia a fármacos. En el cáncer gástrico se ha observado una desregulación de varios miARN tanto en tejido tumoral como en circulación, lo que permite considerarlos como posibles biomarcadores para diagnóstico, pronóstico y seguimiento terapéutico [74], [75], [76], [77].

Diversas investigaciones han identificado patrones conservados de microARN alterados en el cáncer gástrico. Específicamente, se ha reportado un aumento en los niveles de MIR135B-5p, MIR196B-5p y MIR92A-5p en tejido tumoral, mientras que MIR143-3p, MIR204-5p y

MIR133-3p muestran una reducción significativa. La restauración de MIR143-3p en modelos celulares y animales ha demostrado disminuir la proliferación tumoral y aumentar la sensibilidad al cisplatino, lo que indica su función supresora y potencial como objetivo terapéutico. Asimismo, la sobreexpresión de BRD2 (regulada por MIR143-3p) se ha vinculado con un pronóstico menos favorable [78], [79]. El desequilibrio en la selección de brazos de miARN como miR-574-5p y miR-574-3p favorece la progresión tumoral. Un mayor cociente miR-574-5p/-3p se asocia con estadios avanzados y mal pronóstico, sugiriendo posibles aplicaciones terapéuticas al modificar este equilibrio [80], [81].

Modelos pronósticos basados en la expresión de miARN (por ejemplo, hsa-miR-379-3p, hsa-miR-2681-3p, hsa-miR-6499-5p, hsa-miR-6807-3p) permiten estratificar el riesgo y predecir la supervivencia en adenocarcinoma gástrico, con implicaciones en la personalización del tratamiento [76].

3.1.3. Desafíos y tratamiento de datos oncológicos multimodales

La integración de datos multimodales combinando transcriptómica, imágenes médicas e historial clínico ofrece una visión más completa del cáncer de estómago, permitiendo una comprensión más profunda de su biología, progresión y respuesta al tratamiento [82]. Sin embargo, la promesa de la oncología multimodal no está exenta de desafíos considerables. La integración, la estandarización y la interpretación de conjuntos de datos dispares requieren metodologías avanzadas y una infraestructura computacional robusta [83]. Además, la privacidad de los datos, la falta de repositorios compartidos y la necesidad de herramientas analíticas especializadas complican aún más el panorama [84]. Superar estos desafíos es crucial para desbloquear todo el potencial de la oncología de precisión y transformar la atención al paciente con cáncer.

3.1.3.1. Datos faltantes en la información clínica

Uno de los desafíos más persistentes en el análisis de supervivencia con datos clínicos estructurados es la presencia de valores faltantes, un fenómeno que compromete severamente la validez, la generalización y la robustez de los modelos de predicción estadística [85]. En estudios oncológicos de grandes cohortes, los datos clínicos provienen de múltiples instituciones con estándares dispares de recolección, lo que introduce inconsistencias sistemáticas y pérdida parcial de información estructurada [86].

Naturaleza No Aleatoria de los Datos Faltantes

La literatura especializada ha demostrado consistentemente que la falta de datos rara vez es completamente aleatoria. Los patrones de datos faltantes a menudo se alinean con mecanismos de pérdida de datos en los cuales la probabilidad de ausencia de un dato depende de variables observadas o no observadas [87].

En contextos clínicos, los valores faltantes pueden derivarse de múltiples causas: pérdida de seguimiento del paciente, errores manuales en el registro, diferencias institucionales en los protocolos de recolección o incluso el estado clínico del paciente, que puede restringir la captura de ciertos datos (por ejemplo, en pacientes en estado crítico). Este fenómeno es particularmente restrictivo al intentar integrar múltiples tipos de datos. La ausencia de variables clínicas clave puede imposibilitar el entrenamiento de modelos multimodales, reducir el tamaño efectivo de la muestra y deteriorar significativamente el desempeño de los modelos predictivos [87].

Tratamiento de datos faltantes

Para tarar el problema de datos faltantes, existen métodos estadísticos clásicos que comprenden técnicas simples como la imputación por media, mediana o moda, junto con enfoques basados en regresión. Entre las metodologías más robustas destaca la imputación múltiple, particularmente el algoritmo Multiple Imputation by Chained Equations (MICE), ampliamente empleado en estudios biomédicos y epidemiológicos [86], [88]. Este método genera múltiples versiones completas del conjunto de datos mediante modelos condicionales iterativos, reflejando la incertidumbre asociada a los valores faltantes y permitiendo inferencias más válidas que los métodos de imputación única. Paralelamente, han surgido enfoques más avanzados, como la imputación bayesiana, modelos basados en aprendizaje profundo y arquitecturas diseñadas para manejar entradas incompletas, ampliando el repertorio de técnicas disponibles según el contexto analítico y el tipo de dato involucrado [89].

La selección del método de imputación depende de forma crucial del mecanismo de pérdida de datos, el cual se clasifica según la taxonomía de Rubin en Missing Completely at Random (MCAR), Missing at Random (MAR) y Missing Not at Random (MNAR). Cada categoría implica supuestos distintos y determina qué estrategias pueden emplearse sin comprometer la validez estadística y predictiva del modelo. En machine learning, la identificación adecuada del mecanismo de faltantes es esencial para evitar sesgos sistemáticos, garantizar la estabilidad del entrenamiento y preservar la interpretabilidad de los modelos resultantes [90].

Límite máximo aceptado de datos faltantes

En la literatura reciente sobre modelamiento predictivo se observa una tendencia consolidada a utilizar inicialmente criterios operativos para determinar la aceptabilidad de los datos faltantes por variable. La exclusión de predictores con más del 25% de ausencia se ha convertido en una práctica común, dado que proporciones superiores incrementan el riesgo de imputaciones poco confiables y sesgos sistemáticos que comprometen la estabilidad del modelo [91], [92]. Este umbral no debe interpretarse como un criterio absoluto, sino como un estándar pragmático que busca balancear la preservación de información relevante con la garantía de estimaciones estables. Bajo este enfoque, las

variables con ≤25% de valores faltantes son sometidas a imputación, mientras que aquellas que superan dicho límite suelen descartarse, salvo que su valor clínico sea excepcional y justifique un tratamiento diferenciado [93].

Estudio estructural de los patrones de ausencia

El análisis de valores faltantes requiere, antes de cualquier prueba estadística o imputación, comprender cómo se estructuran las ausencias dentro del conjunto de datos. Esta exploración preliminar conocida como estudio estructural de los patrones de ausencia permite identificar si los faltantes se distribuyen aleatoriamente o si aparecen en bloques de variables o grupos de pacientes, lo cual puede introducir sesgos si no se considera adecuadamente [94], [95].

Una estrategia común consiste en transformar cada variable en un indicador binario de ausencia (presente = 0, ausente = 1) y examinar sus patrones de asociación. Las tablas de contingencia permiten cuantificar la co-ocurrencia de faltantes entre pares de variables; sobre ellas pueden calcularse medidas como el coeficiente φ para asociaciones binarias y Cramér's V cuando intervienen variables con múltiples categorías, las cuales permiten evaluar si la ausencia sigue un patrón sistemático o aleatorio [96], [97]. En tablas 2×2, el odds ratio (OR) complementa esta interpretación mediante una medida clara de la fuerza de asociación entre dos eventos de ausencia [98], [99]. La significancia estadística de estos patrones se evalúa mediante la prueba chi-cuadrado o, cuando existen frecuencias esperadas bajas, mediante la prueba exacta de Fisher, que ofrece mayor robustez en muestras pequeñas [94], [100], [101], [102], [103].

El uso de representaciones visuales ha cobrado relevancia en el diagnóstico de patrones de datos faltantes. Herramientas como matrices de calor de ausencia, dendrogramas jerárquicos, mapas de correlación y diagramas UpSet permiten identificar rápidamente bloques de variables que comparten mecanismos de ausencia, casos con registros particularmente incompletos y combinaciones frecuentes de faltantes. Estas visualizaciones se consideran actualmente parte esencial del flujo de trabajo en el análisis de datos incompletos, ya que ayudan a elegir estrategias de imputación coherentes con la estructura observada y a planificar análisis de sensibilidad en contextos clínicos [104], [105], [106].

Identificación de mecanismos de ausencia de datos y sus implicaciones

Posterior a la exploración de los datos faltantes, la comprensión del mecanismo subyacente a los datos faltantes es esencial para seleccionar la estrategia de tratamiento adecuada. En el ámbito clínico, los valores completamente al azar (MCAR) son poco frecuentes; la mayoría de los escenarios corresponden a datos faltantes al azar (MAR), donde la ausencia se explica por otras variables observadas. Por ejemplo, la falta de registro del grado histológico puede asociarse al centro hospitalario o al grupo etario del paciente. En otros casos, la ausencia depende de información no observada, configurando un escenario de datos faltantes no al

azar (MNAR), como cuando determinados estudios diagnósticos sólo se realizan en pacientes con peor estado funcional [107], [108]. El reconocimiento de estos mecanismos es determinante, ya que la imputación múltiple ofrece buenos resultados bajo MAR, pero requiere análisis de sensibilidad adicionales cuando la ausencia es MNAR [96].

La identificación del mecanismo de ausencia no se limita a un ejercicio teórico, sino que requiere procedimientos estadísticos que orienten la decisión metodológica. Una primera estrategia consiste en el análisis exploratorio de los patrones de ausencia, utilizando indicadores binarios de completitud por variable y examinando su relación con otras covariables observadas. Diferencias sistemáticas en estas asociaciones sugieren que el mecanismo difícilmente es MCAR y que corresponde más a MAR o MNAR [109].

Una aproximación complementaria es el ajuste de modelos logísticos para predecir la probabilidad de que un valor esté ausente en función de otras covariables observadas. Si surgen asociaciones estadísticamente significativas, se rechaza la hipótesis de MCAR y se infiere un mecanismo más plausible de MAR [110].

De manera formal, la prueba MCAR de Little [81] constituye el método estadístico más utilizado para evaluar si los datos faltan completamente al azar. Esta prueba plantea como hipótesis nula (H_0) que los datos son MCAR, es decir, que la probabilidad de ausencia es independiente tanto de los valores observados como de los no observados. La hipótesis alterna (H_1) establece que los datos no son MCAR, por lo que el mecanismo corresponde a MAR o MNAR [97], [111].

El procedimiento se basa en comparar las medias de las variables observadas a través de los diferentes patrones de ausencia. Bajo MCAR, estas medias deberían ser estadísticamente equivalentes, dado que la ausencia no depende de la información contenida en los datos. La prueba calcula un estadístico de contraste que, bajo la hipótesis nula, sigue una distribución aproximada χ^2 [111].

La interpretación de los resultados se fundamenta en tres componentes:

- ***Estadístico Chi-cuadrado (X^2):*** refleja la discrepancia global entre las medias observadas por patrón y la media global bajo MCAR. Valores elevados indican diferencias sistemáticas entre patrones, sugiriendo que los datos no son MCAR.
- **Grados de libertad (df):** dependen del número de patrones de ausencia y del número de variables incluidas en el análisis. Representan las comparaciones efectivas realizadas entre las medias de los distintos patrones y la media global.
- **p-value:** cuantifica la probabilidad de obtener un valor de χ^2 igual o mayor al observado si realmente los datos fueran MCAR.

Un p-value alto ($p > 0.05$) implica que no se rechaza la hipótesis nula, lo que respalda el supuesto de MCAR y permite aplicar métodos que lo asumen sin riesgo grave de sesgo.

Un p-value bajo ($p \leq 0.05$) lleva a rechazar la hipótesis nula, indicando que los datos no son MCAR, por lo que el investigador debe asumir mecanismos MAR o MNAR y aplicar imputación múltiple o análisis de sensibilidad [97], [111], [112].

En síntesis, la prueba de Little se convierte en una herramienta diagnóstica: aunque no distingue entre MAR y MNAR, sí permite descartar de manera robusta el supuesto de MCAR. Este diagnóstico previo es fundamental en estudios clínicos, ya que determina si es válido utilizar métodos sencillos o si se requiere el uso de técnicas avanzadas de imputación y modelado [97], [111].

Imputación múltiple por ecuaciones encadenadas (MICE) en investigación clínica

El tratamiento de los datos faltantes mediante la Imputación Múltiple por Ecuaciones Encadenadas (MICE, por sus siglas en inglés), también denominada Fully Conditional Specification (FCS), se ha convertido en una práctica estándar en investigación clínica y epidemiológica debido a su flexibilidad y robustez. A diferencia de los métodos tradicionales, MICE no asume una distribución conjunta específica de todas las variables, sino que modela de forma condicional cada variable incompleta en función de las demás, lo que lo hace especialmente adecuado para bases de datos con variables heterogéneas (continuas, binarias, ordinales o nominales) [113], [114].

El procedimiento de MICE se desarrolla de manera iterativa en cuatro fases:

- **Inicialización:** los valores faltantes se completan con imputaciones preliminares (medias, medianas o valores aleatorios) para permitir el arranque del algoritmo.
- **Ciclo de imputación por variable:** cada variable con datos faltantes se modela condicionalmente usando todas las demás como predictores. Se aplican modelos apropiados: regresión lineal para variables continuas, regresión logística para binarias, regresión multinomial para categóricas y regresión logística ordinal para variables con orden.
- **Iteración encadenada:** el proceso se repite de forma cíclica a través de todas las variables con valores ausentes, produciendo imputaciones sucesivamente más estables.
- **Generación de múltiples datasets:** tras un número suficiente de iteraciones, se obtiene un conjunto de m bases de datos completas. Posteriormente, los análisis se realizan en cada base y se combinan los resultados mediante las reglas de Rubin, lo que asegura que la incertidumbre derivada de la imputación se refleje en los intervalos de confianza y pruebas de hipótesis [115].

Aplicaciones y ventajas de la imputación múltiple

Una característica clave de MICE es la posibilidad de incorporar Predictive Mean Matching (PMM) en el caso de variables continuas. Este método selecciona valores “donantes”

observados con medias pronosticadas similares a las de los casos faltantes, lo que garantiza imputaciones plausibles y conserva la distribución original de los datos, evitando valores fuera de rango [116]. Además, MICE permite la inclusión de interacciones y términos no lineales en los modelos de imputación, lo que lo hace especialmente valioso en estudios clínicos, donde la relación entre variables rara vez es lineal.

3.1.3.2. Alta dimensionalidad en datos genéticos

La integración de datos genéticos en modelos predictivos introduce el desafío de la alta dimensionalidad, caracterizado por un número muy elevado de características moleculares frente a cohortes clínicas relativamente pequeñas. En estudios transcriptómicos de cáncer gástrico es habitual trabajar con miles de miRNAs o genes en muestras que rara vez superan algunos cientos de pacientes, lo cual incrementa el riesgo de sobreajuste, reduce la estabilidad de los modelos y dificulta la reproducibilidad [117], [118]. Este desbalance ha sido ampliamente documentado en la literatura ómica y constituye una limitación central en el desarrollo de modelos de supervivencia basados en datos de expresión.

Dada esta complejidad, los análisis bioinformáticos suelen apoyarse en estrategias de reducción de dimensionalidad y selección preliminar de características, cuyo objetivo es identificar marcadores con cambios de expresión relevantes y un soporte estadístico adecuado, antes de incorporarlos a modelos más complejos. En este contexto, herramientas de visualización como el gráfico de volcán se han consolidado como un recurso estándar para resumir de forma simultánea la magnitud del cambio de expresión y su significancia estadística, permitiendo destacar los genes o miRNAs con mayor evidencia de alteración entre grupos clínicos [119], [120], [121], [122]. Su utilidad radica en ofrecer una visión rápida e interpretable de patrones diferenciales en estudios de alto rendimiento como RNA-seq o perfiles multi-ómicos.

El empleo de estas herramientas no sustituye los análisis estadísticos formales, pero facilita la priorización inicial de biomarcadores y la interpretación general de los datos en escenarios donde el volumen de información supera con amplitud la capacidad de revisión manual. Así, la literatura recomienda su uso como parte de un flujo de trabajo exploratorio que contribuye a manejar la complejidad inherente a los datos ómicos y a respaldar decisiones posteriores en los modelos predictivos [120], [121], [122].

3.1.3.3. Desafíos y tratamiento de imágenes histopatológicas

La integración de imágenes histopatológicas teñidas con hematoxilina y eosina (H&E) en modelos de predicción ha permitido capturar información morfológica rica y clínicamente relevante para el diagnóstico, la estratificación de riesgo y la predicción de supervivencia en cáncer [123], [124], [125], [126]. Sin embargo, el uso de Whole Slide Images (WSI) plantea retos específicos: tamaño en gigapíxeles, heterogeneidad en la adquisición, variabilidad

entre centros y necesidad de anotaciones expertas, todos ellos con impacto directo en la robustez y generalización de los modelos [124], [125].

Desde el punto de vista computacional, el tamaño de las WSI obliga a trabajar a partir de niveles de magnificación estandarizados y a segmentar la lámina en parches de tamaño fijo, como se ilustra en la Fig. 1 y Fig. 2. La normalización del nivel de zoom hacia resoluciones intermedias (por ejemplo, alrededor de 20×) permite equilibrar contexto morfológico, detalle celular y costo computacional, y es una práctica consolidada en patología digital y en estudios de predicción de supervivencia [127], [128], [129], [130], [131], [132], [133]. A partir de esa resolución, los parches constituyen la unidad básica de análisis para arquitecturas basadas en deep learning y enfoques tipo Multiple Instance Learning [134], [135], [136], [137].

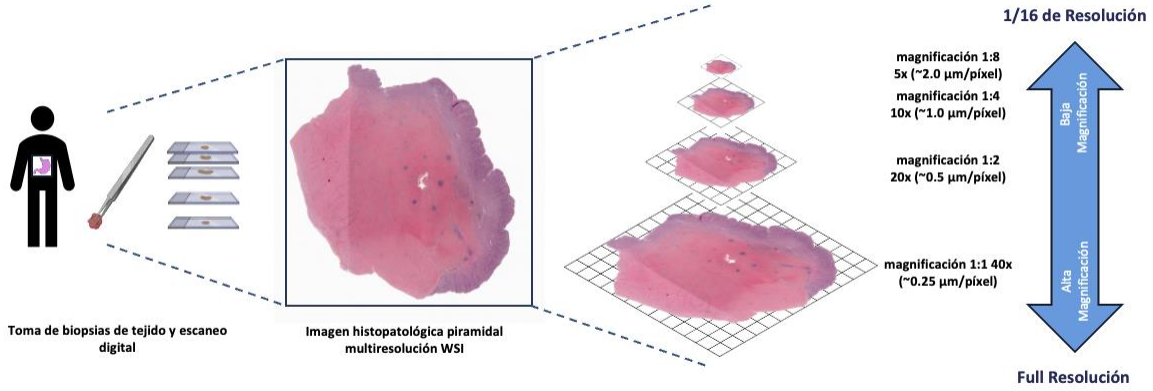


Fig. 1. Estructura piramidal de multirresolución en imágenes histopatológicas digitales (WSI)

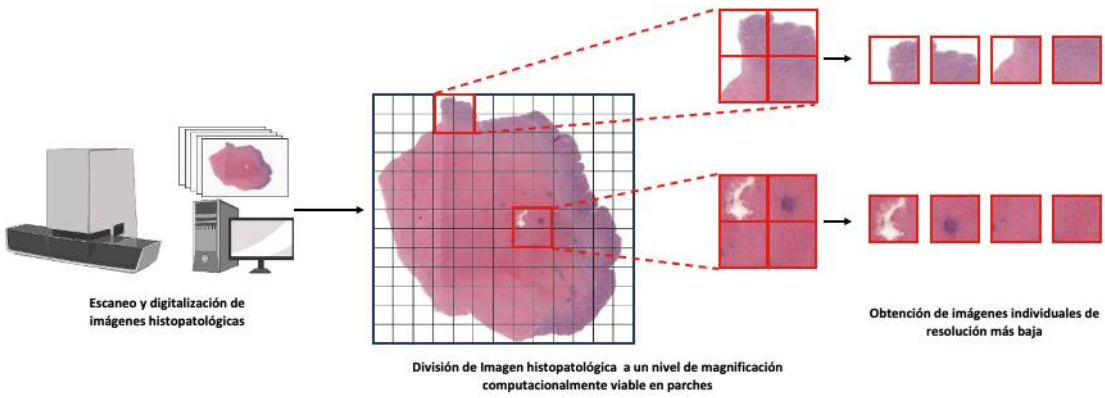


Fig. 2. Extracción de parches de imágenes histopatológicas

Un segundo desafío clave es la variabilidad de color entre laboratorios, escáneres y protocolos de tinción, que puede introducir sesgos y dificultar la transferencia de modelos

entre cohortes. La literatura ha mostrado que la normalización y deconvolución de color mejoran la consistencia cromática y la estabilidad de los modelos, especialmente en H&E [129], [138], [139], [140], [141], [142], [143], [144]. Métodos como Ruifrok–Johnston, Macenko y Vahadane resumidos en la Tabla VI y esquematizados en la Fig. 3 representan enfoques progresivamente más sofisticados para separar los aportes de hematoxilina y eosina y homogeneizar la apariencia de las imágenes, preservando a la vez la arquitectura tisular [140], [142], [143], [145].

Tabla VI. Comparación de métodos de normalización de color basados en deconvolución

Método	Principio	Ventajas	Desventajas
Ruifrok–Johnston	Utiliza vectores de tinción fijos predefinidos (basados en absorción conocida de H&E).	- Implementación simple. - Resultados rápidos y reproducibles. - Adecuado para entornos controlados.	- No se adapta a variaciones entre laboratorios. - Menor fidelidad cromática cuando hay heterogeneidad en tinción.
Macenko	Estima vectores de tinción automáticamente usando análisis estadístico (distribución de colores en OD).	- Se adapta a variabilidad entre muestras. - Amplio uso y validación en literatura. - Preserva comparabilidad entre laboratorios.	- Sensible al ruido y artefactos. - Puede introducir inestabilidad en imágenes muy heterogéneas.
Vahadane	Descomposición mediante factorización no negativa dispersa (Sparse NMF) para separar colorantes y preservar estructura.	- Mejor preservación de la morfología. - Alta fidelidad cromática. - Robusto en cohortes multicéntricas.	- Mayor complejidad computacional. - Requiere mayor tiempo de procesamiento. - Implementación más sofisticada.

Fuente: Elaboración propia usando las siguientes referencias [140], [142], [143], [144], [145]

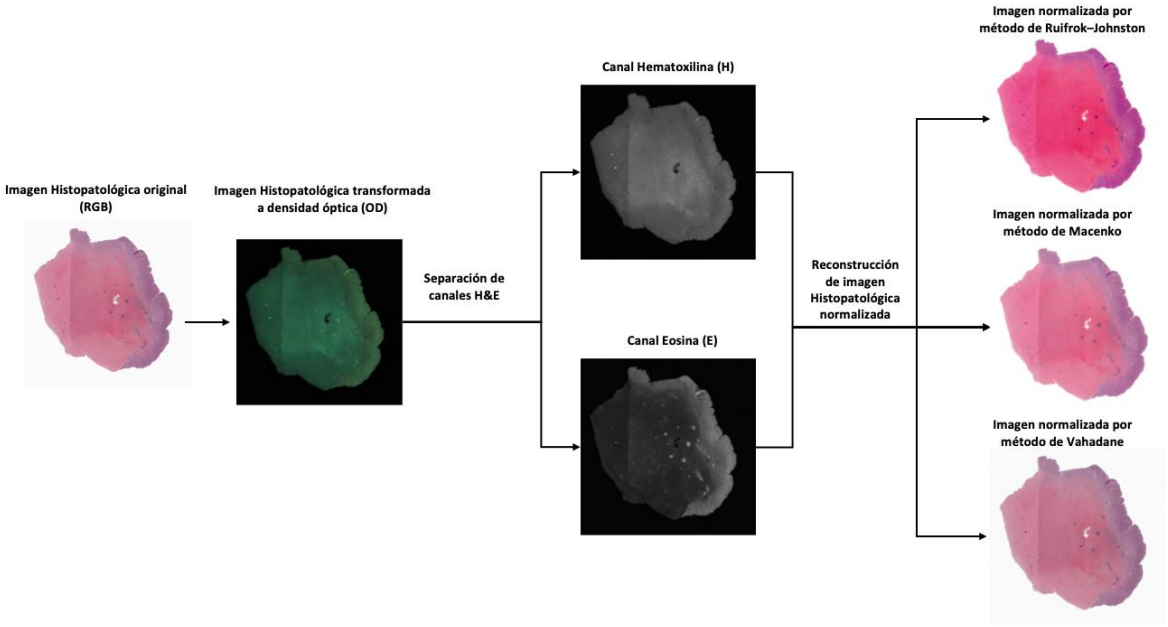


Fig. 3. Proceso de deconvolución de color en imágenes histopatológicas

Por último, no todo el contenido de una WSI es útil para el modelado. Zonas extensas de fondo, vidrio o artefactos de escaneo añaden ruido y consumo computacional sin aportar información biológica. Por ello, es habitual aplicar estrategias de filtrado de parches según contenido tisular, descartando aquellos con baja proporción de tejido, como se muestra en la Fig. 4. Estudios recientes señalan que eliminar entre un 30 % y 50 % de parches sin tejido o redundantes puede mejorar la eficiencia del entrenamiento y concentrar el análisis en regiones relevantes para la predicción [146], [147], [148].

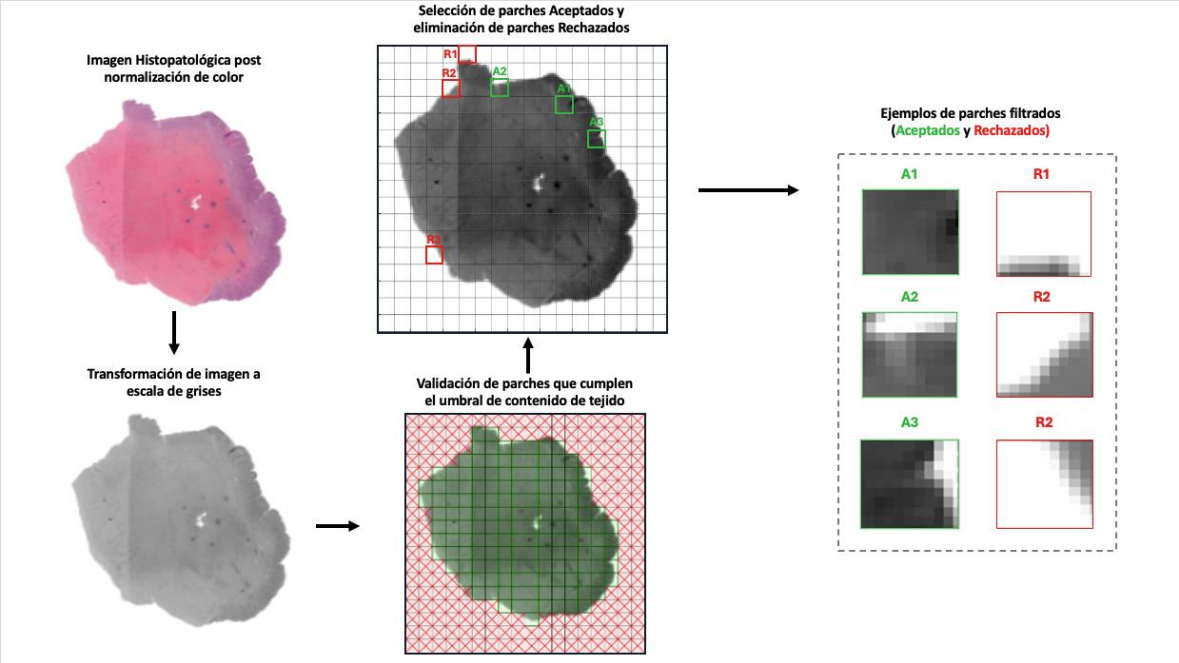


Fig. 4. Proceso de filtrado de parches según contenido tisular en imágenes histopatológicas digitales

En conjunto, estas etapas normalización de magnificación, extracción y filtrado de parches, y normalización de color constituyen el núcleo de los flujos de preprocesamiento en patología digital moderna, y son condición previa para que los modelos de aprendizaje profundo puedan explotar de forma fiable el potencial pronóstico de las imágenes histopatológicas [123], [124], [125], [126], [132], [135], [138], [142], [147], [149].

3.1.4. Selección de características multimodales

La selección de características multimodales constituye una disciplina activa en el análisis de datos biomédicos y aprendizaje automático, aplicable a la integración de datos multiómicos, imágenes médicas y biometría. Los métodos vigentes se agrupan en diversas categorías, cada una con enfoques y utilidades específicas:

3.1.4.1. Selección de características clínicas

La selección de características clínicas constituye un componente fundamental en la construcción de modelos predictivos de supervivencia, particularmente en cáncer gástrico, donde múltiples factores demográficos, anatomopatológicos y tumorales influyen en el pronóstico. Para evaluar el efecto independiente de cada variable sobre el riesgo de muerte, se utiliza el modelo de regresión de Cox, una metodología semiparamétrica ampliamente adoptada en el campo biomédico por su capacidad para manejar datos censurados sin imponer una forma específica a la función de riesgo basal. Este modelo,

introducido por Cox en 1972 [150], permite estimar coeficientes β y analizar el impacto de cada covariable mediante hazard ratios, lo que facilita identificar predictores clínicos relevantes mediante un análisis univariado.

La transformación ordinal de variables categóricas es una práctica estándar que permite preservar relaciones de orden clínicamente significativas, especialmente en características como estadio tumoral, grado de diferenciación, localización anatómica, respuesta terapéutica y tipo histológico. Este último ocupa un lugar destacado desde la clasificación propuesta por Laurén, que establece los subtipos intestinal y difuso, con comportamientos biológicos y pronósticos claramente diferenciados [151]. A esto se suma la evidencia epidemiológica contemporánea, que señala que variables como edad, sexo, estadio patológico, compromiso metastásico y factores anatomopatológicos asociados continúan siendo predictores clave de la supervivencia en cáncer gástrico [9].

El uso de métricas como el coeficiente β, el hazard ratio, el p-valor y el índice de concordancia (C-index) permite evaluar simultáneamente la magnitud, la dirección, la significancia estadística y la capacidad discriminativa de cada característica clínica. El C-index se reconoce como una medida robusta para estimar la capacidad predictiva de modelos de supervivencia en presencia de censura y es ampliamente empleado para comparar el rendimiento de distintos modelos. Organizar las variables de acuerdo con su significancia estadística y su capacidad discriminativa se considera una práctica aceptada en análisis de supervivencia, ya que permite priorizar aquellas características que poseen mayor relevancia clínica y contribución predictiva, en coherencia con los factores pronósticos establecidos en la literatura.

3.1.4.2. Selección de características genéticas

La identificación de biomarcadores genéticos asociados al pronóstico en cáncer gástrico se fundamenta en el análisis de expresión de microARNs (miRNAs), moléculas reguladoras que modulan rutas oncogénicas y se reconocen como indicadores robustos de progresión tumoral, metástasis y supervivencia. Diversos estudios muestran que perfiles específicos de expresión de miRNAs permiten distinguir subgrupos de pacientes con diferente pronóstico y, por tanto, funcionan como biomarcadores de supervivencia y progresión tumoral [152], [153]. En este contexto, la selección de características genéticas no se limita a identificar miRNAs diferencialmente expresados, sino que integra de forma explícita la información de supervivencia mediante modelos de regresión de Cox, con el fin de priorizar aquellos miRNAs cuya expresión se asocia de manera consistente con el desenlace clínico.

Una estrategia ampliamente utilizada consiste en combinar primero un modelo clínico de Cox multivariado con las variables pronósticas más relevantes para obtener un índice de riesgo continuo y estratificar la cohorte en grupos de alto y bajo riesgo, y posteriormente evaluar qué miRNAs presentan diferencias de expresión entre dichos grupos. Este enfoque se alinea con trabajos que construyen firmas pronósticas a partir de datos de miRNA-seq,

en los que la selección inicial de candidatos se basa en el contraste entre perfiles de pacientes con peor y mejor evolución, seguido de análisis de supervivencia mediante regresión de Cox univariada o multivariada [153], [154]. De forma análoga, la comparación de la expresión de miRNAs entre estadios clínicos tempranos (I–II) y avanzados (III–IV) permite identificar moléculas asociadas a la progresión tumoral, coherente con la evidencia de que determinados miRNAs se relacionan con etapas más invasivas y con mayor riesgo de metástasis [152], [155].

Sobre los miRNAs previamente filtrados por expresión diferencial, la aplicación de modelos de regresión de Cox univariados permite cuantificar la asociación individual de cada miRNA con la supervivencia global, estimando hazard ratios y p-valores, lo que facilita descartar aquellos candidatos sin efecto pronóstico consistente. Este esquema que comprende expresión diferencial entre grupos clínicamente definidos, seguida de análisis de supervivencia, se observa en múltiples estudios que construyen firmas de miRNAs para predecir el pronóstico en cáncer gástrico: algunos identifican conjuntos pequeños con valor independiente para predecir la supervivencia [153], [154], mientras que otros emplean cribados de todo el genoma sobre cohortes de TCGA para derivar paneles pronósticos de mayor tamaño [155]. En conjunto, este enfoque permite priorizar miRNAs asociados a riesgo clínico y estadio tumoral, y luego se conservan únicamente aquellos que muestran una asociación estadísticamente relevante con la supervivencia, conformando un conjunto final de biomarcadores genéticos con soporte tanto biológico como estadístico.

3.1.4.3. Selección de características histopatológicas

En el análisis histopatológico del cáncer gástrico, las guías internacionales de reporte estructurado destacan que la interpretación diagnóstica y pronóstica continúa apoyándose en patrones morfológicos visibles en cortes teñidos con hematoxilina y eosina (H&E). Entre ellos se incluyen la arquitectura glandular, la proporción tumor–estroma, la reacción desmoplásica, la presencia de necrosis, la ulceración y los cambios inflamatorios asociados. Estos elementos constituyen criterios centrales en el Histopathology Reporting Guide del International Collaboration on Cancer Reporting (ICCR), que detalla su relevancia para la clasificación y el pronóstico del carcinoma gástrico [156]. Las revisiones contemporáneas sobre análisis digital del cáncer gástrico refuerzan que estas entidades morfológicas siguen siendo esenciales, tanto para el diagnóstico rutinario como para la estratificación del riesgo clínico [157], [158].

Paralelamente, los sistemas de aprendizaje profundo han demostrado alto rendimiento en la identificación de cáncer gástrico a partir de láminas completas [159]; sin embargo, múltiples revisiones subrayan su limitada interpretabilidad, la sensibilidad a variaciones de tinción y la dificultad para traducir sus decisiones en criterios histológicos clásicos [157], [160]. Debido a ello ha emergido un enfoque intermedio: la cuantificación de características morfológicas explícitas, directamente vinculadas al conocimiento histopatológico consolidado (glándulas, estroma, mucina, necrosis, patrones inflamatorios), que pueden ser

integradas en modelos de supervivencia con trazabilidad clínica[157], [160], [161]. Este enfoque combina la robustez cuantitativa del análisis digital con la interpretabilidad requerida por la práctica patológica.

Segmentación del tejido basada en color: respaldo empírico y fundamento cromático

La identificación del tejido sobre el portaobjetos es un paso crítico para cualquier análisis histológico digital. Sin embargo, revisiones sistemáticas demuestran que esta tarea puede resolverse eficazmente utilizando reglas cromáticas simples, sin necesidad de modelos de deep learning. En un análisis exhaustivo de miles de láminas de cohortes TCGA, Ceachi et al. concluyeron que basta con umbralizar la saturación y la luminosidad para separar tejido del fondo blanco con alta precisión [148].

El fundamento radica en las propiedades fotométricas del H&E:

- El fondo carece casi por completo de saturación y presenta valores máximos de brillo (≈255).
- El tejido teñido mantiene una saturación claramente superior y un brillo intermedio debido a la absorción diferencial de hematoxilina y eosina [162]

Valores típicos descritos en la literatura sitúan la saturación del tejido por encima de 30–60 y los valores de brillo por debajo de 230–240, lo que coincide con los rangos necesarios para excluir el fondo blanco sin perder estructuras histológicas relevantes [148], [162]. Esto concuerda con los principios de análisis textural introducidos por Haralick, que exigen eliminar regiones no informativas para obtener mediciones consistentes [163].

Morfometría epitelial y arquitectura tumoral

El epitelio neoplásico es la estructura central del adenocarcinoma gástrico. Las guías de reporte distinguen con claridad entre patrones arquitectónicos propios del subtipo intestinal que implica glándulas más definidas, y del subtipo difuso que se caracteriza por células sueltas y cordones irregulares, ambos con implicaciones pronósticas [156], [158]. Las métricas derivadas del epitelio, como el número de islotes, su área, circularidad, solidez y elongación, reflejan propiedades biológicas de cohesión, polaridad y organización.

La literatura confirma que la complejidad del epitelio tumoral se asocia con agresividad y peor pronóstico. Estudios recientes indican que estructuras fragmentadas o irregulares reflejan un mayor potencial infiltrativo, mientras que contornos redondeados y compactos sugieren grados más bajos de desdiferenciación [157], [161], [164]. Dado que el epitelio teñido con H&E presenta saturación intermedia-alta y brillo menor que el fondo, resulta coherente delimitarlo mediante rangos cromáticos consistentes con estos patrones [162].

Arquitectura glandular y lumen tumoral

La formación glandular es un criterio mayor en la clasificación del adenocarcinoma gástrico. Los tumores del subtipo intestinal suelen mostrar glándulas más organizadas, mientras que la pérdida de estructura glandular indica desdiferenciación y peor pronóstico [156], [158] Las revisiones en análisis digital señalan que la caracterización de la luz glandular (número, forma, proporción de área luminal) aporta información valiosa para capturar patrones de diferenciación y heterogeneidad tumoral [157].

En cortes H&E, las luces glandulares aparecen como regiones de alto brillo y baja saturación, debido a la ausencia relativa de citoplasma y a la menor tinción nuclear. Aunque parte de la evidencia proviene de estudios en otros adenocarcinomas, la lógica fotométrica es común a los tumores glandulares teñidos con H&E [165], [166] Estos patrones permiten delimitar lúmenes y cuantificar su integridad estructural, relevante para la interpretación del grado tumoral y la infiltración [157], [161].

Mucina extracelular y su relevancia pronóstica

El adenocarcinoma gástrico con diferenciación mucinosa presenta lagos de mucina extracelular, cuyo comportamiento clínico difiere del adenocarcinoma no mucinoso. Meng et al. demostraron que este subtipo puede asociarse con peor pronóstico y requiere una caracterización precisa de la cantidad y morfología de la mucina [161].

En H&E, la mucina extracelular destaca por su luminosidad muy alta y baja saturación, lo que la diferencia del epitelio tumoral y del estroma circundante. La cuantificación de su área, tamaño de agregados y proporción de zonas con baja tinción nuclear proporciona información estructural clave para describir el subtipo mucinoso y sus variaciones [158], [161].

Necrosis tumoral y microambiente inflamatorio

La necrosis es uno de los predictores morfológicos más relevantes en el cáncer gástrico. Koskenniemi et al. mostraron que la necrosis tumoral se asocia de forma independiente con disminución de la supervivencia, mayor profundidad invasiva y peor comportamiento biológico [167]. Su aspecto en H&E normalmente visualizado como muy alta luminosidad, escasa tinción nuclear y bordes desestructurados, facilita su identificación digital.

Además, la respuesta inflamatoria perinecrótica constituye un marcador adicional del microambiente tumoral. Regiones con infiltrado linfocitario y macrófagos intensamente teñidos con hematoxilina pueden cuantificarse como densidad de núcleos alrededor de la necrosis, reflejando fenómenos de reparación y respuesta inmunitaria [157], [160], [168].

Estroma, tumor–stroma ratio y reacción desmoplásica

El estroma juega un papel fundamental en la progresión del cáncer gástrico. Kemi et al. demostraron que un tumor–stroma ratio (TSR) elevado se asocia con peor supervivencia y puede utilizarse como factor pronóstico independiente [169]. La integración de métricas que cuantifiquen la proporción de estroma en relación con el tumor permite evaluar la rigidez del microambiente, la respuesta fibroblástica y la interacción tumoral con la matriz extracelular.

La reacción desmoplásica, caracterizada por proliferación de fibroblastos y depósito de colágeno, es especialmente relevante en tumores de patrón infiltrativo. Estudios recientes indican que su intensidad y distribución se correlacionan con invasividad y variación en la supervivencia [158], [164]. La cuantificación del estroma peritumoral en bandas adyacentes al frente invasivo proporciona una estimación robusta de estos fenómenos.

Complejidad geométrica del frente invasor

El frente de invasión tumoral refleja directamente la agresividad del cáncer gástrico. La irregularidad del borde, la fragmentación del epitelio y la presencia de pequeñas agrupaciones celulares (budding) constituyen indicadores bien establecidos de peor pronóstico [158], [164].

El empleo de mediciones multiescala, como la rugosidad del borde y la dimensión fractal, se fundamenta en el análisis geométrico clásico aplicado a imágenes biomédicas. Haralick y otros autores han demostrado que la complejidad estructural en múltiples escalas revela patrones de organización tisular relevantes que no se capturan en mediciones uniescalares [163]. Estas métricas reflejan la pérdida de cohesión, la variabilidad del patrón infiltrativo y la heterogeneidad estructural del tumor.

Textura tisular como indicador de heterogeneidad tumoral

La textura es una propiedad esencial en la caracterización digital de tumores. Las características texturales propuestas por Haralick, como la entropía y energía permiten capturar patrones de variabilidad nuclear, densidad estromal y organización tisular [163]. Las revisiones recientes indican que los tumores más agresivos y con peor pronóstico tienden a mostrar mayor entropía y menor uniformidad, lo que refleja heterogeneidad celular y desorganización arquitectónica [157], [158], [164].

Estas características complementan las mediciones morfológicas clásicas y ofrecen una representación cuantitativa de la complejidad tisular que puede aportar valor en modelos de supervivencia.

Ulceración tumoral y cambios superficiales

La ulceración se considera un indicador de enfermedad avanzada en el cáncer gástrico. Se ha descrito que los tumores ulcerados suelen presentar necrosis extensa, infiltración profunda y un microambiente inflamatorio más intenso, lo cual se asocia con peor pronóstico [170]. En cortes H&E, combina áreas extremadamente claras (fibrina superficial, detritos) con zonas muy oscuras (hemorragia, necrosis profunda), además de una marcada reducción de tejido epitelial viable. Su reconocimiento digital permite estimar de manera aproximada la presencia de estos patrones complejos.

3.1.5. Modelos para la predicción de la supervivencia

El análisis de supervivencia se centra en una variable de tiempo hasta evento, usualmente denotada por T, que representa el tiempo transcurrido hasta que ocurre un evento de interés (por ejemplo, muerte, recurrencia o progresión de la enfermedad). El objetivo principal es describir y modelar la función de supervivencia, definida como:

$$S(t) = P(T > t),$$

es decir, la probabilidad de que un paciente permanezca vivo (o libre del evento) más allá del tiempo t. De forma complementaria, se define la función de riesgo $h(t)$ como la tasa instantánea a la que ocurre el evento en el tiempo t, condicionada a haber sobrevivido hasta ese momento. Matemáticamente,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

Ambas funciones están relacionadas mediante la función de riesgo acumulado $H(t)$, dada por la integral de $h(t)$ a lo largo del tiempo,

$$H(t) = \int_0^t h(u)du,$$

De forma que la supervivencia puede escribirse como

$$S(t) = \exp\big(-H(t)\big).$$

Esta relación es fundamental: una vez se dispone de una estimación del riesgo (instantáneo o acumulado), es posible obtener la probabilidad de supervivencia en cualquier tiempo t a través de la función exponencial [171].

A partir de estos conceptos básicos, se han desarrollado diversos modelos para estimar $S(t)$ y estudiar el efecto de covariables clínicas y biológicas. Entre los más utilizados se encuentran el estimador de Kaplan–Meier, el modelo de Cox de riesgos proporcionales y, más recientemente, los métodos de aprendizaje automático como el Random Survival Forest, que extienden estas ideas a contextos no lineales y de alta dimensión [171], [172].

Estimador de Kaplan–Meier

El estimador de Kaplan–Meier es un método no paramétrico clásico para estimar la función de supervivencia sin asumir ninguna forma específica para la distribución de T. Fue introducido por Kaplan y Meier en 1958 y se diseñó explícitamente para tratar datos censurados, es decir, casos en los que no se observa el evento durante el periodo de seguimiento [173].

La idea central es construir la supervivencia total como un producto de probabilidades condicionales. Sea

$$t_{(1)} < t_{(2)} < \cdots < t_{(k)}$$

la secuencia ordenada de tiempos en los que se observan eventos (por ejemplo, fallecimientos). Para cada tiempo $t_{(j)}$, se define:

- n_j: número de pacientes en riesgo justo antes de $t_{(j)}$ (es decir, que todavía no han presentado el evento ni han sido censurados antes de ese tiempo);
- d_j: número de eventos que ocurren exactamente en $t_{(j)}$.

La probabilidad condicional de sobrevivir más allá de $t_{(j)}$, dado que se estaba vivo justo antes, se estima como

$$\widehat{p_j} = 1 - \frac{d_j}{n_j}.$$

El estimador de Kaplan–Meier de la función de supervivencia hasta un tiempo t se obtiene multiplicando todas estas probabilidades condicionales para los tiempos de evento menores o iguales que t:

$$\hat{S}(t) = \prod_{t_{(j)} \le t} \left(1 - \frac{d_j}{n_j}\right).$$

Desde un punto de vista intuitivo, cada factor $(1 - d_j/n_j)$ representa la probabilidad de "superar" el intervalo alrededor del tiempo $t_{(j)}$ sin que ocurra el evento. El producto acumulado de estos factores refleja la probabilidad de que el paciente sobreviva a todos los

intervalos sucesivos hasta el tiempo t. La censura se incorpora actualizando correctamente n_j: cuando un paciente es censurado, se retira del conjunto en riesgo a partir del tiempo de censura, pero no se contabiliza como evento.

El estimador de Kaplan–Meier permite obtener curvas de supervivencia escalonadas y comparar grupos (por ejemplo, distintos estadios tumorales o estrategias terapéuticas) mediante pruebas como el log-rank. Sin embargo, este enfoque es principalmente descriptivo: no incluye explícitamente covariables a nivel individual, por lo que no permite cuantificar directamente el efecto de características clínicas o moleculares sobre el riesgo [171], [173]

Modelo de Cox de riesgos proporcionales

El modelo de Cox de riesgos proporcionales, propuesto por David Cox en 1972, constituye uno de los pilares del análisis de supervivencia moderno. Este modelo permite estudiar cómo un conjunto de covariables influye sobre el riesgo de presentar el evento, manteniendo una formulación relativamente flexible al no imponer una forma paramétrica específica para el riesgo base.

El modelo de Cox de riesgos proporcionales, propuesto por David Cox en 1972, constituye uno de los pilares del análisis de supervivencia moderno. Este modelo permite estudiar cómo un conjunto de covariables $X = (x_1, x_2, \dots, x_p)$ influye sobre el riesgo de presentar el evento, manteniendo una formulación relativamente flexible al no imponer una forma paramétrica específica para el riesgo base [150].

En el modelo de Cox, la función de riesgo condicional se expresa como

$$h(t|X) = h_0(t)\exp(\beta^T X),$$

donde:

- $h_0(t)$ es la función de riesgo base (o “baseline”), común a todos los individuos;
- β es el vector de coeficientes que cuantifica el efecto de cada covariable sobre el riesgo;
- $\exp(\beta^T X)$ actúa como un factor multiplicativo que ajusta el riesgo base según el perfil del paciente.

La suposición clave es la de riesgos proporcionales: el cociente de riesgos entre dos individuos con covariables x_1 y x_2 es

$$\frac{h(t|X_1)}{h(t|X_2)} = \exp(\beta^T(X_1 - X_2)),$$

lo que implica que, esta razón de riesgos es constante en el tiempo (no depende de t). Esta propiedad permite interpretar $\exp(\beta_j)$ como un hazard ratio: el cambio relativo en el riesgo asociado a un aumento unitario de la covariable x_j, manteniendo las demás constantes.

Aunque el modelo se formula en términos de riesgo, también permite obtener la probabilidad de supervivencia. Partiendo de la relación

$$H(t|X) = \int_0^t h(u|X)du = \exp(\beta^T X) \int_0^t h_0(u)du = \exp(\beta^T X) H_0(t),$$

donde $H_0(t)$ es el riesgo acumulado base, la función de supervivencia condicional se escribe como

$$S(t|X) = \exp\big(-H(t|X)\big) = \exp\big(-\exp(\beta X)\, H_0(t)\big).$$

A menudo se reexpresa esta relación como

$$S(t|X) = [S_0(t)]^{\exp(\beta X)},$$

donde $S_0(t) = \exp(-H_0(t))$ es la supervivencia base. De esta forma, la probabilidad de supervivencia para un paciente concreto se obtiene elevando la curva de supervivencia base a una potencia que depende exponencialmente de sus covariables.

Un aspecto importante del modelo de Cox es que los coeficientes β se estiman mediante la verosimilitud parcial, la cual se construye únicamente a partir del orden de los tiempos de evento y de los conjuntos en riesgo, sin necesidad de especificar $h_0(t)$. Posteriormente, el riesgo base y la supervivencia base pueden estimarse de forma no paramétrica [150].

El modelo de Cox se ha utilizado extensamente en investigación clínica y oncológica para identificar factores pronósticos y construir modelos de riesgo, gracias a su equilibrio entre interpretabilidad (a través de los hazard ratios) y flexibilidad al no fijar una forma funcional para el riesgo en el tiempo [171].

Random Survival Forest

El Random Survival Forest (RSF) representa una extensión de los bosques aleatorios al ámbito del análisis de supervivencia. Fue introducido por Ishwaran y colaboradores como un método no paramétrico basado en ensambles de árboles para manejar datos censurados, relaciones no lineales y posibles interacciones complejas entre covariables [174].

Al igual que en los bosques aleatorios clásicos, el RSF construye un gran número de árboles de decisión, cada uno entrenado sobre una muestra bootstrap de los datos. En cada nodo,

en lugar de utilizar criterios estándar como la reducción de la impureza en clasificación o la varianza en regresión, se emplean criterios específicos de supervivencia, por ejemplo, estadísticas basadas en pruebas log-rank para maximizar la diferencia de supervivencia entre las ramas hijas [174].

El funcionamiento puede entenderse en tres pasos conceptuales:

- **Construcción de árboles de supervivencia**
  Cada árbol se entrena con una muestra bootstrap y, en cada nodo, se selecciona aleatoriamente un subconjunto de covariables para buscar la mejor partición según un criterio de supervivencia (por ejemplo, la mayor separación entre curvas de supervivencia de los grupos resultantes). El crecimiento continúa hasta que se cumplen criterios de parada (como un tamaño mínimo de nodo).

- **Estimación de supervivencia en los nodos terminales**
  En cada nodo terminal, se dispone de un conjunto de individuos con sus tiempos y estados (evento o censura). Para este subconjunto se estima una curva de supervivencia, típicamente mediante el estimador de Kaplan–Meier. Cada árbol asocia así, a cualquier observación que caiga en un nodo terminal, una función de supervivencia $\hat{S}_b(t|X)$, donde b indica el árbol [174].

- **Promedio de ensamble**
  La supervivencia del bosque para un individuo con covariables X se obtiene promediando las curvas de supervivencia aportadas por todos los árboles:

$$\hat{S}_{RSF}(t|X) = \frac{1}{B}\sum_{b=1}^{B} \hat{S}_b(t|X),$$

  donde B es el número total de árboles. De nuevo, este resultado puede interpretarse como una probabilidad estimada de sobrevivir más allá de t, pero ahora incorporando de forma flexible relaciones no lineales, interacciones entre variables y estructuras complejas en los datos.

El RSF ofrece varias ventajas frente a los modelos clásicos: maneja de forma natural grandes conjuntos de variables, puede modelar efectos no lineales y no depende de la suposición de riesgos proporcionales. Además, proporciona medidas de importancia de variables que ayudan a identificar los predictores más influyentes en el riesgo. Sin embargo, su interpretación no es tan directa como en el modelo de Cox, ya que no se centra en coeficientes ni hazard ratios, sino en la estructura global del bosque y en predicciones de supervivencia individuales [172], [175].

Estudios comparativos han mostrado que, en escenarios con relaciones altamente no lineales o cuando las suposiciones del modelo de Cox no se cumplen plenamente, el RSF

puede ofrecer un mejor desempeño predictivo, manteniendo la capacidad de manejar censura y tiempos hasta evento [172], [176].

DeepSurv: red neuronal tipo Cox

En los últimos años, el uso de redes neuronales profundas se ha extendido al análisis de supervivencia con el objetivo de capturar relaciones no lineales y complejas entre las covariables y el riesgo de evento, especialmente en contextos con datos de alta dimensión (por ejemplo, genómica e imágenes médicas) [177], [178].

La idea general de estos modelos es mantener el marco probabilístico del análisis de supervivencia clásico (funciones de riesgo, riesgo acumulado y supervivencia), pero reemplazando la combinación lineal de covariables por una red neuronal profunda capaz de aprender representaciones complejas de los datos. De esta forma, se busca mejorar la capacidad predictiva manteniendo la coherencia con la teoría estadística subyacente.

DeepSurv es uno de los modelos más representativos de esta familia. Fue propuesto por Katzman et al. como una red neuronal profunda que extiende el modelo de Cox de riesgos proporcionales, sustituyendo el predictor lineal $\beta^T X$ por una función no lineal $f_\theta(X)$ aprendida por una red neuronal [179].

En el modelo de Cox clásico, el riesgo condicional se escribe como:

$$h(t|X) = h_0(t)\ \exp(\beta X),$$

mientras que en DeepSurv se plantea:

$$h(t|X) = h_0(t)\ \exp(f_\theta(X)),$$

donde:

- $h_0(t)$ sigue siendo el riesgo base común,
- $f_\theta(X)$ es la salida de la red neuronal (un escalar), que actúa como un score de riesgo no lineal,
- θ representa todos los parámetros, pesos y sesgos de la red.

La estructura típica de DeepSurv consiste en varias capas densas (fully connected) con funciones de activación no lineales (como ReLU), que transforman progresivamente las covariables de entrada en un valor escalar $f_\theta(X)$ [179].

Desde el punto de vista de la supervivencia, la relación entre riesgo y probabilidad de supervivencia se mantiene igual que en el modelo de Cox. El riesgo acumulado condicional es:

$$H(t|X) = \exp(f_\theta(X))\ H_0(t),$$

y, en consecuencia, la función de supervivencia condicional se expresa como:

$$S(t|X) = \exp\big(-H(t|X)\big) = \exp\Big(-\exp\big(f_\theta(X)\big)\, H_0(t)\Big).$$

Equivalente a:

$$S(t|X) = [S_0(t)]^{\exp(f_\theta(X))},$$

Donde $S_0(t) = \exp(-H_0(t))$ es la supervivencia base.

Es decir, la probabilidad de que un paciente sobreviva más allá de un tiempo t se obtiene como una transformación exponencial del riesgo acumulado, pero ahora el factor que ajusta el riesgo (el "score" del paciente) lo provee una red neuronal que puede capturar patrones complejos y no lineales en las covariables.

DeepSurv se entrena maximizando (o, en la práctica, minimizando la negativa de) la verosimilitud parcial de Cox, adaptada para que la salida de la red sea el predictor:

$$\ell(\theta) = \sum_{i \in \mathcal{E}} \left[f_\theta(X_i) - \log\left(\sum_{i \in R_i} \exp\big(f_\theta(X_j)\big) \right) \right],$$

donde,

- $\mathcal{E}$ es el conjunto de individuos que presenta el evento,
- R_i es el conjunto "en riesgo" en el instante del evento del individuo i.

Intuitivamente, el modelo aprende a asignar valores de riesgo mayores a quienes fallecen antes, en comparación con quienes permanecen más tiempo sin evento. Esta función de pérdida se optimiza mediante descenso de gradiente y sus variantes (por ejemplo, Adam), como en otros modelos de deep learning [179].

El resultado final es una red neuronal que conserva la interpretación en términos de riesgo relativo (como en Cox), pero con una frontera de decisión mucho más flexible, capaz de capturar interacciones y efectos no lineales entre covariables clínicas, genéticas o de imagen. Estudios empíricos han mostrado que DeepSurv puede igualar o superar el

desempeño de modelos de supervivencia clásicos en distintos conjuntos de datos clínicos [178], [179].

En aplicaciones reales, DeepSurv produce para cada paciente un score de riesgo $f_\theta(X)$. Para obtener la probabilidad de supervivencia $S(t|X)$, se sigue un esquema similar al del modelo de Cox:

- Se estima primero el riesgo acumulado base $H_0(t)$ y la supervivencia base a $S_0(t)$ partir de los datos de entrenamiento, usando procedimientos no paramétricos análogos a los del modelo de Cox.
- Para un nuevo paciente, se calcula $f_\theta(X)$ con la red neuronal.
- La supervivencia condicional se obtiene como:

$$S(t|X) = [S_0(t)]^{\exp(f_\theta(X))}.$$

Interpretado de forma sencilla: si $\exp(f_\theta(X)) > 1$, el paciente tiene un riesgo mayor que el promedio y su curva de supervivencia decrece más rápido que la curva base; si es menor que 1, la curva decrece más lentamente, indicando mejor pronóstico [179], [180].

3.1.6. Optimización y evaluación de modelos de supervivencia

La calidad de un modelo de supervivencia no depende solo de la elección del algoritmo (Cox, Random Survival Forest, DeepSurv, etc.), sino también de cómo se ajustan sus parámetros, se controla el sobreajuste y se evalúa su capacidad predictiva. Dado que los datos de supervivencia incluyen censura y una dimensión temporal explícita, tanto la optimización como la evaluación requieren adaptaciones específicas frente a los métodos clásicos de regresión o clasificación [181].

3.1.6.1. Parámetros e hiperparámetros de los modelos de supervivencia

Modelo de Cox (clásico y penalizado)

En el modelo de Cox, la forma básica del riesgo es:

$$h(t|X) = h_0(t)\exp(\beta^T X),$$

donde:

- $h_0(t)$: riesgo base, estimado de manera no paramétrica.
- β: coeficientes del modelo. Cada β_j expresa el cambio relativo en el riesgo por unidad de cambio de la covariable x_j.
- $\exp(\beta_j) > 1$: la covariable incrementa el riesgo,
- $\exp(\beta_j) < 1$: la covariable se asocia con menor riesgo.

En variantes penalizadas del modelo de Cox (LASSO, Ridge, Elastic Net) aparecen hiperparámetros adicionales [182]:

- λ (parámetro de penalización): controla la fuerza de la regularización.
  - λ grande: más contracción de coeficientes, menos sobreajuste, pero mayor riesgo de subajuste.
  - λ pequeño: coeficientes más libres, mayor riesgo de sobreajuste.
- α (en Elastic Net): mezcla entre penalización L1 y L2.
  - $\alpha = 1$: LASSO puro (selección de variables).
  - $\alpha = 0$: Ridge puro (contracción sin "apagar" del todo variables).
  - $0 < \alpha < 1$: combinación de ambos efectos.

Random Survival Forest (RSF)

En los Random Survival Forests, el modelo está definido por un conjunto de árboles de supervivencia. Algunos hiperparámetros relevantes son [174]:

- ***Número de árboles (ntree):*** Cantidad de árboles en el bosque.
  - Más árboles implican menor varianza, pero más coste computacional.
  - Suele fijarse en cientos o miles.
- ***Número de variables candidatas por división (mtry):*** Número de covariables que se consideran al buscar la mejor partición en cada nodo.
  - Valor bajo implica mayor aleatoriedad, más diversidad entre árboles.
  - Valor alto implica divisiones más "óptimas" por árbol, pero menos diversidad.
- ***Tamaño mínimo de nodo terminal (nodesize):*** Número mínimo de observaciones en un nodo para dejar de dividir.
  - Nodesize pequeño genera árboles más profundos, mayor flexibilidad, mayor riesgo de sobreajuste.
  - Nodesize grande genera árboles más "lisos", menor varianza.
- ***Profundidad máxima del árbol (maxdepth, si se usa):*** Límite en el número de niveles por árbol.
  - Limitar la profundidad es otra forma de controlar complejidad.
- ***Criterio de división (splitrule):*** Regla para evaluar la calidad de una partición, por ejemplo, estadístico de log-rank o criterios basados en funciones de riesgo acumulado.
- ***Fracción de muestra por árbol (sample_fraction / bootstrap):*** Proporción de datos utilizada para construir cada árbol (muestra bootstrap).
  - Controla la variabilidad de las estimaciones out-of-bag (OOB).

Además, el RSF incorpora medidas de importancia de variables (VIMP) y criterios de profundidad mínima para selección de características.

DeepSurv y Redes Neuronales de Supervivencia

En DeepSurv, la red neuronal reemplaza el predictor lineal de Cox por una función no lineal $f_\theta(X)$. Los hiperparámetros típicos son [179]:

- ***Arquitectura de la red***
  - Número de capas ocultas.
  - Número de neuronas por capa.
  - Tipo de activación (ReLU, tanh, etc.).
- ***Hiperparámetros de entrenamiento***
  - Learning rate: tamaño del paso en el descenso de gradiente.
  - Batch size: número de muestras por actualización de parámetros.
  - Número de épocas (epochs): cuántas veces se recorre el conjunto de entrenamiento.
- ***Regularización***
  - L2 / weight decay: penalización sobre la magnitud de los pesos para evitar coeficientes grandes.
  - Dropout rate: proporción de neuronas "apagadas" aleatoriamente durante el entrenamiento para reducir coadaptaciones.
  - Early stopping: detener el entrenamiento cuando la métrica de validación deja de mejorar.

Estos hiperparámetros se ajustan para maximizar la verosimilitud parcial de Cox o variaciones de esta, utilizando validación cruzada o un conjunto de validación separado.

3.1.6.2. Validación cruzada en modelos de supervivencia

La validación cruzada (VC) permite estimar cómo se comportará el modelo en datos nuevos. En supervivencia, debe adaptarse a la presencia de censura y tiempos de evento.

k-fold estratificado por evento y censura

En el esquema k-fold:

- Se divide la cohorte en k subconjuntos (folds).
- En cada iteración, uno se usa como validación y los restantes como entrenamiento.
- Se ajusta el modelo en el conjunto de entrenamiento y se evalúa con una métrica para datos censurados (por ejemplo, C-index o Brier score dependiente del tiempo)

[181].

En supervivencia se suele:

- Estratificar por evento: asegurar proporciones similares de eventos y censura en cada fold.
- Evitar particiones que generen folds sin suficientes eventos, ya que dificultan la estimación de curvas de supervivencia y penalizan la estabilidad del C-index o del Brier score.

Validación cruzada anidada

Cuando se hace búsqueda de hiperparámetros, es recomendable usar VC anidada:

- ***Bucle interno:*** elegir hiperparámetros (por ejemplo, λ en Cox penalizado, ntree/mtry/nodesize en RSF, o learning rate y número de capas en DeepSurv).
- Bucle externo: evaluar el rendimiento del modelo ya optimizado, evitando que el conjunto de validación interno "contamine" la estimación de desempeño [179].

Este enfoque reduce el optimismo en la estimación del rendimiento, algo especialmente relevante cuando se comparan varios algoritmos de supervivencia.

Bootstrap y error out-of-bag

En RSF es muy frecuente usar el error out-of-bag (OOB) como medida de validación interna: cada árbol se entrena con una muestra bootstrap y las observaciones no utilizadas para ese árbol (OOB) se usan para evaluar predicciones. El promedio sobre todos los árboles proporciona una estimación casi "gratuita" del error generalizado, sin necesidad de una VC explícita [174].

De forma más general, los métodos bootstrap .632 y .632+ permiten combinar el error dentro y fuera de la muestra para obtener estimaciones menos sesgadas del rendimiento de modelos de supervivencia [174].

3.1.6.3. Regularización y prevención del sobreajuste

El sobreajuste en supervivencia se traduce en modelos que predicen muy bien los datos de entrenamiento, pero fallan al generalizar a nuevas cohortes. Esto es especialmente crítico en contextos con muchas covariables y tamaños muestrales moderados.

Modelos de Cox penalizados

La incorporación de penalizaciones L1, L2 o Elastic Net en el modelo de Cox permite:

- Reducir la varianza de los coeficientes, evitando estimaciones inestables.
- Seleccionar variables (en el caso de LASSO y Elastic Net).

En términos prácticos:

- La penalización L1 tiende a llevar algunos coeficientes exactamente a cero, generando un modelo más parsimonioso.
- La penalización L2 reduce la magnitud de todos los coeficientes, pero sin anularlos.
- Elastic Net combina ambas, útil cuando hay grupos de variables correlacionadas.

El valor óptimo de λ suele escogerse mediante validación cruzada, eligiendo el que maximiza el C-index o minimiza el Brier score en los folds .

Control de complejidad en Random Survival Forest

En RSF, el sobreajuste se controla ajustando hiperparámetros estructurales:

- ***nodesize y maxdepth:*** limitar el crecimiento de los árboles evita nodos con muy pocos individuos (que ajustan ruido).
- ***mtry:*** elegir un número moderado de variables por división favorece la diversidad del bosque y reduce la dependencia excesiva de unas pocas covariables.
- **ntree:** aumentar el número de árboles reduce la varianza, aunque con rendimientos decrecientes a partir de cierto punto.

Además, el uso de importancia de variables (VIMP) y profundidad mínima permite identificar y eventualmente descartar variables poco informativas, reduciendo complejidad del modelo [174].

Regularización en DeepSurv y redes neuronales

En modelos profundos de supervivencia, las estrategias de regularización típicas incluyen:

- ***Weight decay (L2):*** penaliza pesos grandes en la función de pérdida, favoreciendo soluciones más suaves.
- **Dropout:** durante el entrenamiento se “apagan” conexiones de forma aleatoria con una cierta probabilidad (dropout rate), obligando a la red a no depender de rutas específicas.
- ***Early stopping:*** se monitoriza el C-index o el Brier score en un conjunto de validación y se detiene el entrenamiento cuando deja de mejorar.
- ***Reducción de la arquitectura:*** limitar el número de capas y neuronas para evitar que el modelo tenga capacidad excesiva frente al tamaño de muestra.

La elección de estos hiperparámetros se hace típicamente mediante validación cruzada, priorizando configuraciones que ofrezcan buen equilibrio entre discriminación y robustez.

3.1.6.4. Métricas de evaluación y herramientas gráficas

La evaluación de modelos de supervivencia combina métricas de discriminación, calibración y bondad de ajuste global. Ninguna métrica es suficiente por sí sola; se recomienda emplear varias de manera complementaria [181].

Índice de concordancia (C-index)

El C-index mide la capacidad del modelo para ordenar correctamente los pacientes según su riesgo: es la proporción de pares de individuos en los que la predicción de riesgo coincide con el orden real de los tiempos de evento, considerando adecuadamente la censura [183].

Interpretación:

- 0.5: rendimiento equivalente al azar.
- 0.7: buena capacidad discriminativa en muchos contextos clínicos.

Brier Score y Brier Score integrado

El Brier Score dependiente del tiempo cuantifica el error cuadrático entre la supervivencia predicha y el estado observado (evento o no) en un tiempo específico t, corrigiendo por censura mediante ponderaciones de probabilidad de censura inversa.

- Valores bajos indican mejor calibración y precisión.
- El Integrated Brier Score (IBS) resume el error medio a lo largo de un intervalo de tiempo.

Curvas de calibración para supervivencia

La calibración evalúa si las probabilidades de supervivencia predichas coinciden con las observadas en cada grupo de riesgo .
Procedimiento habitual:

- Agrupar pacientes en cuartiles o quintiles según la probabilidad de supervivencia predicha en un tiempo t^*.
- Para cada grupo, estimar la supervivencia observada (por ejemplo, con Kaplan–Meier).
- Representar en un gráfico la probabilidad predicha (eje X) frente a la observada (eje Y).

Como se observa en la Fig. 5 la línea diagonal (45°) representa calibración perfecta. Los puntos o curvas de cada grupo se comparan con esta referencia. Una desviación sistemática por encima o por debajo indica sobreestimación o subestimación del riesgo por parte del modelo.

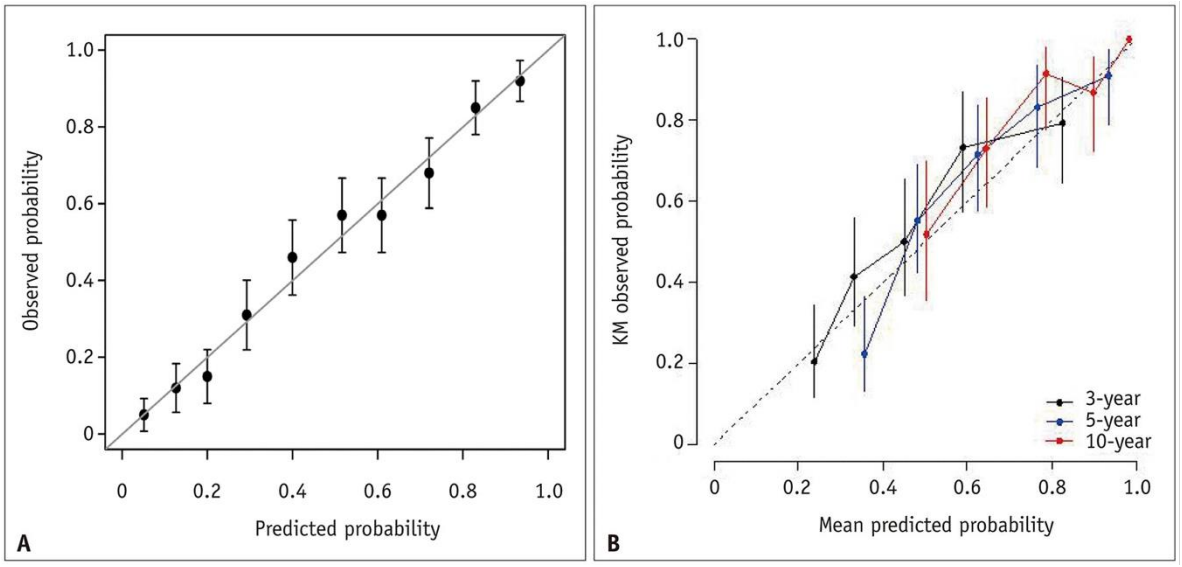


Fig. 5. Ejemplo de gráfico de curvas de calibración [181]

Curvas ROC y AUC dependientes del tiempo

Las curvas ROC dependientes del tiempo extienden el concepto clásico de sensibilidad y especificidad a contextos donde el estado del paciente cambia con el tiempo (por ejemplo, vivo vs muerto a los 12 o 36 meses) [181], [183].

Para un tiempo t^*:

- ***Sensibilidad(t*):*** proporción de pacientes que han presentado el evento antes de t^* y que el modelo identifica como de alto riesgo.
- **Especificidad(t*):** proporción de pacientes que no han presentado el evento antes de t^* y que el modelo clasifica correctamente como de bajo riesgo.

Como se observa en la Fig. 6, la curva *ROC(t*)* representa sensibilidad frente a 1 – especificidad para distintos umbrales del score de riesgo. El *AUC(t*)* resume el área bajo la curva y mide la discriminación en ese tiempo. Es frecuente mostrar varias curvas ROC (por ejemplo, para 1, 3 y 5 años) en el mismo gráfico o en paneles separados.

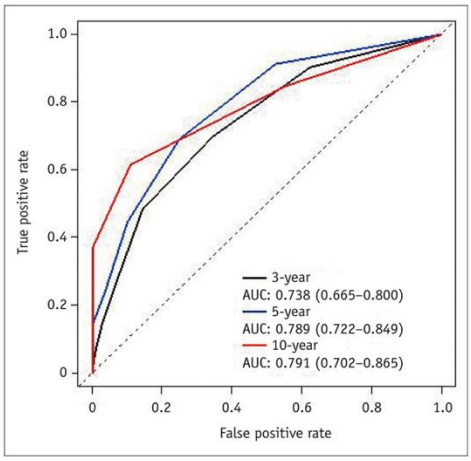


Fig. 6. Ejemplo de gráfico de curvas ROC y AUC dependientes del tiempo [181]

Curvas de supervivencia por grupos de riesgo

Otra forma intuitiva de evaluar un modelo es estratificar a los pacientes según su score de riesgo (por ejemplo, bajo, intermedio y alto) y estimar para cada grupo una curva de supervivencia (Kaplan–Meier) [174], [181].

Como se observa en la Fig. 7, en el eje X se representa el tiempo y en el eje Y la probabilidad de supervivencia. Se trazan tres curvas escalonadas, una por grupo de riesgo. Un modelo útil debe producir curvas claramente separadas y ordenadas de acuerdo con el riesgo (la de alto riesgo por debajo, la de bajo riesgo por arriba).

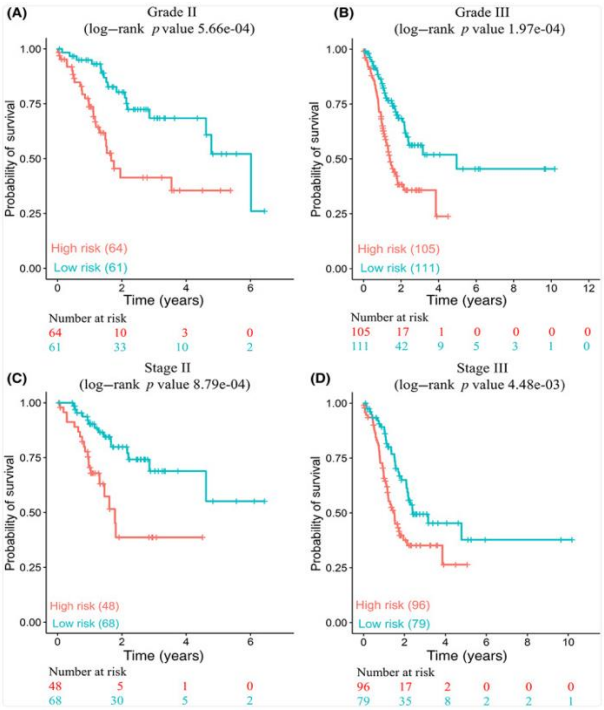


Fig. 7. Ejemplo de curvas de supervivencia por grupo de riesgo y grado [184]

4. METODOLOGÍA

El desarrollo metodológico de esta investigación se alinea con el modelo CRISP-DM (Cross-Industry Standard Process for Data Mining), adaptado al análisis multimodal del cáncer gástrico. Este marco proporciona una estructura clara y secuencial para la preparación, integración y explotación de datos clínicos, genéticos e histopatológicos, permitiendo abordar de manera ordenada los desafíos inherentes al procesamiento de grandes volúmenes de información heterogénea. En este estudio, CRISP-DM se adopta como guía para organizar las fases analíticas que van desde la comprensión y la caracterización de los datos, hasta su depuración, transformación, modelado y evaluación, asegurando coherencia metodológica, reproducibilidad y una alineación explícita entre los objetivos del proyecto y las decisiones técnicas implementadas.

4.1. Comprensión del dominio y requerimientos de datos

La primera fase se orientó a establecer los requerimientos conceptuales y analíticos necesarios para el desarrollo del modelo de predicción de supervivencia. Para ello, se revisaron los elementos clínicos, genéticos e histopatológicos que intervienen en la progresión del cáncer gástrico y en las métricas de supervivencia, con el fin de determinar qué tipos de variables, formatos y niveles de granularidad son indispensables para el análisis. En esta fase se definieron las características mínimas que debian cumplir los repositorios de datos a emplear, incluyendo: disponibilidad de variables clínicas relevantes para modelos de supervivencia, presencia de identificadores únicos que permitieran la integración multimodal, acceso a perfiles genómicos procesables y disponibilidad de imágenes histopatológicas en formatos adecuados para análisis computacional.

Asimismo, se estableció el procedimiento sistemático para la identificación y selección de repositorios biomédicos: revisión de bases de datos especializadas, análisis de las modalidades de información ofrecidas, evaluación de la calidad y estandarización de los metadatos, verificación de las condiciones éticas de uso y determinación del nivel de accesibilidad requerido para su descarga y procesamiento. Esta fase permitió delimitar los requisitos de los conjuntos de datos a utilizar y establecer las condiciones necesarias para avanzar hacia la adquisición, preparación e integración de la información multimodal.

4.2. Adquisición y selección de datos

La adquisición de información se realizó a partir de repositorios biomédicos internacionales que contienen datos clínicos, genómicos e imágenes histopatológicas asociados a cáncer gástrico. Para ello se implementó un proceso sistemático orientado a identificar fuentes confiables, estandarizadas y con potencial para la integración multimodal requerida por el proyecto.

En una primera etapa, se efectuaron búsquedas estructuradas en bases de datos bibliográficas (PubMed, Web of Science y Scopus) y en registros especializados de datos biomédicos. Estas búsquedas permitieron identificar repositorios que alojan datos abiertos o de acceso controlado relevantes para la investigación, entre ellos plataformas como The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO), ArrayExpress y dbGaP, así como directorios especializados (FAIRsharing, re3data, OpenAIRE). Para ello se emplearon combinaciones de términos relacionados con datos clínicos, genómicos e imagenológicos en cáncer, priorizando aquellos vínculos con cáncer gástrico.

Posteriormente, los repositorios identificados fueron evaluados con base en criterios metodológicos previamente definidos:

- Accesibilidad y disponibilidad de los datos, incluyendo licencias y requisitos de uso.
- Cobertura multimodal, verificando la existencia de información clínica, genética o transcriptómica y, cuando estaba disponible, imágenes histopatológicas.
- Trazabilidad mediante identificadores únicos, que permitieran correlacionar información de distintas modalidades a nivel de paciente.
- Calidad, completitud y estandarización de los metadatos, necesarios para el análisis reproducible.
- Volumen y diversidad de casos, suficiente para el entrenamiento y validación de modelos de supervivencia.
- Actualización y mantenimiento del repositorio, garantizando estabilidad y vigencia de los datos.
- Cumplimiento ético y regulatorio, particularmente en relación con protección de datos sensibles y aprobación de comités de ética.

Este proceso permitió identificar un conjunto de repositorios que cumplían con los estándares necesarios para la preparación, integración y modelado subsecuente. La selección final de las fuentes de datos específicas se presenta detalladamente en el apartado de resultados, junto con sus características y contribución a las etapas posteriores del análisis.

4.3. Comprensión de los datos

Una vez identificadas las fuentes de información pertinentes, se procedió a realizar un análisis exploratorio inicial de cada modalidad de datos con el fin de evaluar su estructura, completitud y pertinencia para los objetivos del estudio. En esta etapa se revisaron los metadatos asociados a los conjuntos clínicos, genéticos e histopatológicos, verificando la consistencia de los identificadores de paciente, la disponibilidad de variables de desenlace (tiempo de supervivencia y estatus vital) y la presencia de medidas demográficas, clínicas y transcriptómicas requeridas para los modelos de supervivencia.

Para los datos clínicos, se examinaron las variables relacionadas con características demográficas, factores tumorales, tratamientos y desenlaces, identificando valores faltantes, codificaciones heterogéneas y posibles inconsistencias. En el caso de los datos genéticos, se revisaron matrices de expresión y anotaciones transcriptómicas, evaluando su dimensionalidad, distribución y la presencia de posibles sesgos derivados de la técnica experimental. En cuanto a las imágenes histopatológicas, se verificó la disponibilidad de metadatos técnicos (tipo de escáner, resolución, nivel de magnificación) y la correspondencia entre cada imagen y el identificador del paciente.

Asimismo, se realizó una primera evaluación de la calidad de los datos, considerando la existencia de valores atípicos, niveles significativos de datos faltantes, duplicados, diferencias en escalas y formatos, así como la viabilidad técnica para su posterior integración multimodal. Este reconocimiento inicial permitió anticipar las transformaciones necesarias en la siguiente fase de preparación, incluyendo imputación, estandarización, normalización, armonización de identificadores y definición de criterios de inclusión/exclusión de pacientes.

Con esta comprensión inicial, se estableció una hoja de ruta para la preparación de los datos clínicos, genómicos e histopatológicos, garantizando un proceso coherente con los requisitos de los modelos de supervivencia a desarrollar en etapas posteriores.

4.4. Preparación de los datos

La preparación de los datos constituyó una fase central del proceso analítico, en la cual se transformaron las distintas modalidades de información (clínica, genética e histopatológica) en estructuras consistentes y adecuadas para su posterior integración y modelado. Siguiendo los lineamientos de la fase Data Preparation de CRISP-DM, se llevaron a cabo procedimientos sistemáticos de limpieza, depuración, estandarización y armonización de todas las fuentes seleccionadas. Estas actividades incluyeron el análisis y tratamiento de valores faltantes en las variables clínicas, la normalización y preprocesamiento de los datos transcriptómicos, y la aplicación de controles de calidad, normalización cromática y segmentación tisular sobre las imágenes histopatológicas. El objetivo fue asegurar un conjunto de datos uniforme, trazable y coherente a nivel de paciente, preservando la integridad y relevancia biológica de la información para su uso posterior en los modelos de predicción de supervivencia.

4.4.1. Preparación de datos clínicos

El conjunto de datos clínicos presentó diferentes proporciones de valores faltantes distribuidos entre variables numéricas y categóricas. Como paso inicial, y siguiendo los criterios de calidad descritos en la sección 3.1.3.1. Datos faltantes en la información clínica

del marco teórico, se eliminaron del análisis aquellas variables con más del 25% de valores faltantes, con el fin de reducir el sesgo y la incertidumbre asociada a la imputación de datos.

Posteriormente, se realizó un análisis estructural de los patrones de ausencia para identificar su naturaleza y las relaciones subyacentes entre variables. Para ello se evaluaron asociaciones bivariadas entre la ausencia de datos y otras características mediante:

- el coeficiente φ (phi) para variables binarias,
- la medida V de Cramér para variables categóricas multiclase,
- el cálculo de odds ratio (OR) para estimar la fuerza de asociación,
- pruebas de chi-cuadrado de Pearson o prueba exacta de Fisher, según las frecuencias observadas.

Este análisis se complementó con herramientas gráficas, incluyendo:

- matriz binaria de completitud,
- matriz de calor de correlación de valores faltantes,
- dendrograma jerárquico basado en similitud de patrones de ausencia,
- diagrama UpSet para visualizar combinaciones de valores faltantes simultáneos.

Con el fin de establecer si la ausencia era aleatoria, se aplicó la prueba MCAR de Little. Los resultados indicaron que los datos no eran MCAR, por lo cual se descartaron métodos simples de imputación. En consecuencia, se empleó el método de Imputación Múltiple por Ecuaciones Encadenadas (MICE), adecuado para preservar las relaciones entre variables clínicas. El proceso permitió recuperar valores con estabilidad en sus distribuciones, comparando los datos imputados con los valores originales para garantizar la coherencia del conjunto final.

4.4.2. Preparación de datos genéticos

La preparación de los datos genéticos se centró en garantizar que la matriz de expresión miRNA-seq utilizada proviniera de un proceso estandarizado y técnicamente consistente. Para ello, se emplearon los archivos procesados del repositorio TCGA, los cuales ya incorporan las etapas fundamentales del pipeline de preprocesamiento: filtrado de lecturas de baja calidad, eliminación de adaptadores, depuración de secuencias de baja complejidad, mapeo contra un genoma de referencia y detección de lecturas duplicadas. Estas operaciones aseguran que las señales de expresión disponibles correspondan únicamente a lecturas válidas y permiten minimizar el ruido técnico asociado a la secuenciación. Los valores de expresión suministrados por TCGA se encuentran expresados en Reads Per Million (RPM), una medida normalizada que corrige diferencias en la profundidad de secuenciación entre muestras.

Adicionalmente, se realizó una verificación estructural del conjunto de datos con el fin de evaluar su completitud, consistencia y condiciones para el análisis posterior. Este proceso incluyó: revisión de la correspondencia entre identificadores de paciente (Case ID) y perfiles de expresión, comprobación de que cada miRNA estuviera correctamente anotado, inspección de la distribución general de los valores de expresión y evaluación preliminar de la variabilidad entre miRNAs. Estas acciones permitieron confirmar que el formato, la estructura y la coherencia del conjunto eran adecuados para su integración con los datos clínicos e histopatológicos, sin aplicar en esta fase ningún procedimiento de filtrado o selección de transcritos, dado que el objetivo era únicamente asegurar la preparación y calidad inicial del dataset.

4.4.3. Preparación de datos histopatológicos

El procesamiento de las imágenes histopatológicas se desarrolló siguiendo los principios expuestos en la sección 3.1.3.3. Desafíos y tratamiento de imágenes histopatológicas del marco teórico, con el propósito de garantizar un conjunto de WSIs homogéneo, libre de artefactos críticos y con suficiente contenido tisular para una extracción confiable de características.

Control de calidad basado en HistoQC

Como primer paso, se aplicó un control de calidad sistemático mediante HistoQC para identificar artefactos como desenfoque, sobreexposición, escaso contenido tisular o degradación de bordes. Cada lámina fue clasificada en tres categorías:

- Ok: condiciones adecuadas de enfoque, brillo y cobertura tisular;
- Review: láminas aceptables pero con advertencias;
- Bad: láminas con defectos severos.

Solo las láminas clasificadas como Ok y Review se conservaron, mientras que las Bad fueron excluidas del análisis. La retención de las etiquetas de calidad permitió evaluar posteriormente el impacto del QC en las características extraídas y en el desempeño de los modelos.

Normalización de color mediante Macenko

Con el fin de reducir la variabilidad no biológica derivada de diferencias en tinción o escaneo, se aplicó la normalización de color mediante el método de Macenko. Para ello:

- se seleccionó una lámina de referencia escaneada a ~20× (0.5 µm/píxel),
- se generó un thumbnail de máximo 8192 px para estimar los vectores de tinción,
- la normalización se aplicó parche por parche utilizando deconvolución óptica.

Esta estrategia permitió homogenizar la apariencia cromática de las láminas, preservando al mismo tiempo su estructura tisular.

Selección de regiones tisulares y extracción de parches

La extracción de parches siguió criterios establecidos en la literatura para balancear resolución diagnóstica, contexto morfológico y viabilidad computacional. Para ello:

- se trabajó a una magnificación equivalente a 20× (~0.5 µm/píxel),
- se empleó un tamaño de parche de 1024×1024 px (≈ 512×512 µm),
- se estableció un traslape del 25% entre parches (stride de 768 px).

Este tamaño permitió capturar simultáneamente arquitectura tumoral, celularidad y continuidad espacial relevante.

Filtrado de parches según contenido tisular

Para garantizar la calidad morfológica del conjunto final, se implementó un criterio estricto de contenido tisular, reteniendo únicamente los parches con ≥ 50% de tejido. El cálculo se realizó en el espacio HSV:

- la saturación (S) permitió separar tejido de fondo,
- el valor (V) permitió descartar zonas excesivamente brillantes.

Este procedimiento, ampliamente reportado en la literatura, redujo la inclusión de parches irrelevantes y mejoró la precisión del modelado posterior.

4.5. Extracción y selección de características

La extracción y selección de características constituyó una etapa crítica dentro del proceso analítico, ya que permitió identificar, dentro de cada una de las modalidades de datos (clínica, genética e histopatológica), aquellas variables con mayor potencial explicativo y capacidad predictiva frente al desenlace de interés: la supervivencia de los pacientes con cáncer gástrico. Este procedimiento se ejecutó de forma independiente para cada modalidad y posteriormente sirvió como fundamento para la integración multimodal expuesta en las fases subsiguientes.

4.5.1. Selección de características clínicas

Con el fin de determinar las variables clínicas más relevantes asociadas con la supervivencia, se aplicó un análisis univariado basado en el modelo de regresión de Cox proporcional de

riesgos, siguiendo los fundamentos teóricos presentados en la sección 3.1.4.1. Selección de características clínicas del marco teórico. Este modelo permitió evaluar el efecto individual de cada variable sobre el riesgo de muerte, controlando simultáneamente el tiempo de seguimiento.

Previamente, las variables categóricas fueron transformadas mediante codificación ordinal para garantizar compatibilidad con los modelos de supervivencia. Se incluyeron variables demográficas y clínicas como *Age at Index*, *Gender*, *Tissue Origin*, *Classification of Tumor*, *Treatment Types*, *Tumor Grade*, *AJCC Stage*, *Residual Disease* y *Disease Response*, utilizando *duration* y *event* como variables de tiempo y censura, respectivamente.

Para cada variable se estimaron algunas de las métricas definidas en la sección 3.1.6.4. Métricas de evaluación y herramientas gráficas:

- β (coeficiente) para interpretar dirección y magnitud del efecto.
- *Hazard Ratio* ($HR = \exp(\beta)$), que indica el cambio relativo en el riesgo por unidad de incremento.
- *p-valor*, evaluando significancia estadística ($\alpha = 0.05$).
- *Índice de concordancia* (*C-index*) como medida de discriminación del modelo.

Las variables fueron ordenadas según su peso estadístico y relevancia clínica documentada en la sección 3.1.2.3. Características clínicas y genéticas relevantes en el análisis de supervivencia, priorizando aquellas con significancia estadística y consistencia con los factores pronósticos ampliamente validados en cáncer gástrico. Este subconjunto enriquecido constituyó el grupo final de características clínicas empleadas en la integración multimodal.

4.5.2. Selección de características genéticas

En primera instancia, se construyó un modelo de regresión de Cox multivariado utilizando las características clínicas seleccionadas previamente. A partir de este modelo se generó un índice de riesgo continuo, el cual permitió estratificar a los pacientes en dos grupos (alto y bajo riesgo) según la mediana del puntaje.

Análisis de expresión diferencial mediante Volcano Plots

El índice de riesgo clínico se vinculó con las matrices miRNA-seq mediante los identificadores estandarizados del TCGA, permitiendo etiquetar cada muestra genética con su respectivo grupo pronóstico. Sobre esta base y aplicando lo expuesto en la sección 3.1.3.2. Alta dimensionalidad en datos genéticos del marco teórico:

- Se generó un primer análisis de expresión génica diferencial comparando pacientes

de alto vs. bajo riesgo.
- Paralelamente, se realizó un segundo análisis contrastando estadios tempranos (I–II) frente a avanzados (III–IV), con el fin de capturar miRNAs relacionados con progresión tumoral.

En ambos casos se identificaron miRNAs diferencialmente expresados empleando criterios estadísticos basados en $Fold\ Change$ y $p - value$.

Evaluación mediante regresión cox

Los miRNAs candidatos resultantes de ambos análisis fueron intersectados para obtener un conjunto robusto de biomarcadores. Posteriormente:

- Cada miRNA fue evaluado mediante un modelo univariado de Cox,
- Conservándose aquellos con significancia estadística o marcada tendencia pronóstica.

Este proceso permitió obtener un conjunto final de miRNAs interpretables y estadísticamente relevantes, adecuados para su integración en modelos de supervivencia.

4.5.3. Extracción y selección de características histopatológicas

Una vez obtenidos los parches tisulares normalizados mediante el procedimiento descrito previamente, se llevó a cabo la extracción cuantitativa de características histopatológicas. El objetivo fue capturar propiedades morfológicas, estructurales y texturales del tumor y su microambiente, de forma trazable e interpretable desde la perspectiva patológica según lo presentado en la sección 3.1.4.3. Selección de características histopatológicas del marco teórico.

La delimitación del tejido y de sus principales componentes (epitelio tumoral, mucina, estroma, necrosis) se realizó mediante reglas fotométricas basadas en las propiedades ópticas de la tinción H&E. Para ello se utilizaron umbrales de saturación y brillo en los espacios de color HSV y LAB, previamente definidos según rangos descritos en la literatura especializada para separar fondo blanco de tejido real y diferenciar patrones histológicos característicos. Los parámetros auxiliares como tamaños mínimos de objetos, kernels para operaciones morfológicas o filtros de ruido, se establecieron antes del análisis estadístico y se mantuvieron dentro de valores estándar reportados en estudios de análisis digital de histopatología, garantizando así estabilidad computacional y ausencia de sesgos derivados de ajustes manuales.

A partir de cada parche tisular se cuantificaron diversas estructuras relevantes para el diagnóstico y el pronóstico del adenocarcinoma gástrico, incluyendo:

- ***Morfometría epitelial:*** número y tamaño de islotes tumorales, circularidad, solidez y elongación, indicadores de cohesión celular, grado de desdiferenciación y complejidad arquitectónica.
- ***Arquitectura glandular:*** presencia y morfología del lumen tumoral, proporción de área luminal y características de forma, asociadas al grado de diferenciación en tumores del subtipo intestinal.
- ***Desmoplasia y estroma peritumoral:*** mediciones dentro de bandas alrededor del frente tumoral, fundamentadas en la relevancia clínico-patológica del microambiente fibroblástico.
- ***Necrosis y respuesta inflamatoria:*** proporción de áreas necróticas y densidad de infiltrado perinecrótico, indicadores de agresividad tumoral y actividad inmunitaria local.
- ***Mucina extracelular:*** fracción de área mucinosa, tamaño de agregados y grado de tinción nuclear residual, relevantes para el subtipo mucinoso.
- ***Complejidad geométrica del frente invasor:*** rugosidad del borde y dimensión fractal calculada mediante análisis multiescala, reflejando la naturaleza infiltrativa del tumor.
- ***Características texturales:*** entropía, uniformidad y gradientes de intensidad, que permiten medir heterogeneidad nuclear y variación estructural.
- ***Medidas complementarias:*** métricas de nitidez, densidad de bordes y proxies de ulceración, útiles para evaluar la calidad local del tejido y patrones superficiales avanzados.

Posteriormente, las características calculadas por parche fueron agregadas a nivel de caso mediante estadísticas robustas (medianas, máximos y percentiles superiores), con el fin de obtener descriptores globales representativos de cada paciente. Este procedimiento permitió construir un conjunto de variables histopatológicas interpretables, cuantitativas y clínicamente justificadas, aptas para su integración en los modelos de supervivencia.

Finalmente, debido a los altos requerimientos computacionales asociados al procesamiento de imágenes en alta resolución y a la extracción paralela de miles de parches por WSI, el pipeline fue ejecutado en una instancia de cómputo de Amazon EC2 con múltiples núcleos, permitiendo optimizar tiempos, garantizar reproducibilidad y manejar eficientemente los volúmenes de datos sin comprometer la calidad del procesamiento.

4.6. Modelado

La fase de modelado consistió en la implementación, ajuste y evaluación inicial de tres enfoques complementarios para la predicción de la supervivencia: un modelo clásico basado en regresión de Cox, un modelo basado en ensambles no paramétricos (Random Survival Forest) y un modelo profundo de supervivencia (DeepSurv) presentados en la

sección 3.1.5. Modelos para la predicción de la supervivencia del marco teórico. El propósito fue comparar métodos con diferentes supuestos, complejidades y capacidades de representación, aprovechando tanto la interpretabilidad clínica del modelo tradicional como la robustez y flexibilidad de los métodos modernos.

Preparación del conjunto de datos para modelado

Antes de entrenar los modelos, se construyó una matriz final de características integrada a nivel de paciente, combinando:

- Variables clínicas seleccionadas (sección 4.5.1),
- Biomarcadores genéticos (miRNAs) seleccionados (sección 4.5.2),
- Características histopatológicas agregadas a nivel de caso (sección 4.5.3).

Las variables de supervivencia incluyeron la duración del seguimiento (*duration*) y el estatus del evento (*event*).

El dataset fue dividido en entrenamiento y prueba mediante una partición estratificada por el evento, con el fin de mantener una proporción equilibrada de casos censurados y no censurados. Todas las características continuas fueron normalizadas y aquellas con distribuciones altamente asimétricas fueron transformadas mediante funciones logarítmicas o rank-based cuando fue necesario para mejorar la estabilidad del entrenamiento.

4.6.1. Modelo de regresión de Cox proporcional de riesgos

El modelo de Cox se empleó como referencia base por su amplia adopción clínica y su interpretabilidad. Se implementó un modelo de Cox penalizado para manejar la alta dimensionalidad y multicolinealidad potencial de las características multimodales.

El proceso incluyó:

- Ajuste del modelo en el conjunto de entrenamiento.
- Selección del parámetro de penalización mediante validación cruzada interna.
- Estimación de los coeficientes β para cada variable y sus hazard ratios asociados.
- Evaluación de la suposición de proporcionalidad de riesgos mediante residuos de Schoenfeld.

La salida del modelo permitió identificar las variables que contribuían más al riesgo relativo de mortalidad, manteniendo un marco interpretativo alineado con la práctica clínica.

4.6.2. Modelo Random Survival Forest

Dado que la regresión de Cox asume proporcionalidad de riesgos y relaciones lineales entre predictores y hazard, se empleó un modelo de Random Survival Forest para capturar relaciones no lineales y posibles interacciones entre características.

El entrenamiento del RSF incluyó:

- Construcción de múltiples árboles de supervivencia sobre muestras bootstrap.
- División de nodos basada en maximización de log-rank o criterios equivalentes.
- Optimización de hiperparámetros como número de árboles, profundidad máxima y tamaño mínimo de nodos terminales.
- Cálculo de importancia de variables mediante medidas de permutación adaptadas a censura.

El RSF permitió evaluar el peso relativo de cada modalidad (clínica, genética, histopatológica) y la presencia de patrones no lineales relevantes para el pronóstico.

4.6.3. Modelo DeepSurv

Para capturar estructuras complejas en datos multimodales, se implementó un modelo DeepSurv, una red neuronal que optimiza directamente la función parcial de verosimilitud del modelo de Cox.
El procedimiento de entrenamiento incluyó:

- Definición de una arquitectura densa con múltiples capas intermedias y funciones de activación no lineales.
- Aplicación de regularización (dropout y penalizaciones L2) para reducir sobreajuste.
- Entrenamiento mediante algoritmos de optimización adaptativa aplicados a la función parcial del riesgo.
- Uso de early stopping basado en la pérdida de validación para asegurar estabilidad.

DeepSurv permitió modelar interacciones complejas entre características y capturar patrones de riesgo no lineales, manteniendo compatibilidad conceptual con el marco de riesgos proporcionales.

4.6.4. Comparación entre modelos y preparación para la fase de evaluación

Finalizado el entrenamiento, cada modelo generó un índice de riesgo (risk score) continuo para cada paciente en el conjunto de prueba. Estos valores fueron utilizados para:

- Evaluar discriminación,
- Comparar desempeño entre métodos,
- Analizar la contribución relativa de las características multimodales,
- Preparar insumos para la etapa de evaluación.

Esta fase permitió disponer de tres aproximaciones metodológicas complementarias:

- un modelo interpretable de referencia (Cox),
- un modelo basado en ensambles robusto a relaciones no lineales (RSF),
- y un modelo profundo orientado a capturar interacciones complejas (DeepSurv).

4.7. Evaluación de modelos

La fase de evaluación tuvo como objetivo determinar el desempeño real de los tres modelos implementados (Cox, Random Survival Forest y DeepSurv) utilizando métricas específicas para datos censurados las cuales se presentan en la sección 3.1.6. Optimización y evaluación de modelos de supervivencia del marco teórico. Esta etapa permitió comparar su capacidad predictiva y verificar su utilidad para la estratificación del riesgo en pacientes con cáncer gástrico.

Métricas empleadas

Se utilizaron cuatro métricas complementarias:

- ***Índice de concordancia (C-index):*** Evaluó la capacidad discriminativa global del modelo, midiendo qué tan bien ordena correctamente los tiempos de supervivencia entre individuos.
- **Integrated Brier Score (IBS):** Cuantificó el error de predicción a lo largo del tiempo, integrando el Brier Score y proporcionando una medida robusta de calibración temporal.
- **ROC dependiente del tiempo y área bajo la curva AUC(t):** Se calcularon curvas ROC específicas para momentos clave del seguimiento (por ejemplo, 1, 3 y 5 años). El AUC(t) permitió evaluar la capacidad del modelo para distinguir entre pacientes que fallecen antes del tiempo *t* versus aquellos que sobreviven más allá de *t*, incorporando adecuadamente la censura. Esta métrica complementa al C-index, pues ofrece discriminación puntual en distintos horizontes temporales relevantes clínicamente.
- **Separación de grupos de riesgo:** Los índices de riesgo derivados de cada modelo se utilizaron para estratificar a los pacientes en grupos (alto y bajo riesgo). Se evaluó la separación entre curvas de Kaplan–Meier, y la significancia mediante log-rank test.

Validación y control de sobreajuste

Se evaluó el desempeño sobre el conjunto de prueba y se aplicaron controles específicos según el modelo:

- Cox: verificación del supuesto de proporcionalidad.
- RSF: estabilidad del ranking de importancia de variables y número óptimo de árboles.
- DeepSurv: inspección de curvas de pérdida entrenamiento/validación y early stopping.

Comparación entre modelos

Los modelos se compararon según:

- discriminación global (*C-index*),
- discriminación puntual (*AUC(t)*),
- calibración (*IBS*),
- y separación clínica de riesgo (*KM + log-rank*).

Esta combinación de métricas permitió identificar el modelo con mejor equilibrio entre precisión, estabilidad y utilidad clínica.

5. ANÁLISIS Y RESULTADOS

5.1. Comprensión de requerimientos de datos

5.2. Adquisición y selección de datos

La adquisición de información se realizó a partir de repositorios biomédicos internacionales que albergan datos clínicos, genómicos y de imágenes histopatológicas relevantes para el cáncer gástrico. Para este fin, se implementó un procedimiento sistemático que combinó búsquedas bibliográficas en bases como PubMed, Web of Science y Scopus, exploración directa de repositorios de datos biomédicos (The Cancer Genome Atlas – TCGA, Gene Expression Omnibus – GEO, ArrayExpress, dbGaP) y consulta de directorios especializados (FAIRsharing, re3data, OpenAIRE), empleando combinaciones de términos relacionados con "gastric cancer," "clinical data", "genomic data", "histopathology images" y "open cancer datasets".

Los repositorios identificados se evaluaron conforme a los criterios metodológicos definidos en la sección 4.2. de metodología: (i) accesibilidad y licenciamiento; (ii) cobertura

multimodal (clínica, genética/transcriptómica, imagenológica); (iii) trazabilidad mediante identificadores únicos a nivel de paciente; (iv) calidad, completitud y estandarización de metadatos; (v) volumen y diversidad de casos; (vi) actualización y mantenimiento; y (vii) cumplimiento ético y regulatorio. Sobre esta base, se elaboró una matriz comparativa como se observa en la Tabla VII entre los repositorios y directorios considerados.

Tabla VII. Comparación de repositorios y directorios según criterios metodológicos

Criterio	TCGA (GDC)	GEO (NCBI)	ArrayExpress (EBI)	dbGaP (NIH)	FAIRsharing	re3data	OpenAIRE
Accesibilidad y tipo de acceso	Acceso abierto a muchos datos + acceso controlado para datos sensibles	Acceso abierto	Acceso abierto	Principalmente acceso controlado	Acceso abierto (directorio)	Acceso abierto (directorio)	Acceso abierto (directorio/agregador)
Cobertura multimodal	Clínica, genómica (miRNA-seq, RNA-seq, etc.) e imágenes histopatológicas	Principalmente datos transcriptómicos y algunos clínicos	Principalmente datos transcriptómicos y algunos clínicos	Datos genéticos con fenotipos asociados	No aplica (no almacena datos)	No aplica (no almacena datos)	No aplica (enlaza a diversos repositorios)
Trazabilidad (IDs únicos por paciente)	Muy alta (identificadores consistentes a nivel de caso/paciente)	Alta (Accession IDs por serie/muestra)	Alta (Accession IDs por experimento/muestra)	Variable, depende del estudio	No aplica	No aplica	No aplica
Calidad y estandarización de metadatos	Alta, con esquemas de anotación definidos y documentación detallada	Variable según el estudio, con plantillas comunes del NCBI	Alta, sigue estándares como MIAME/MINSEQE	Alta, con descripciones estructuradas	Alta, catálogo de estándares FAIR	Alta, catálogo de repositorios	Alta, alineada a políticas de la UE
Volumen y diversidad de casos	Alto número de casos oncológicos, incluyendo cohorte TCGA-STAD	Medio–alto, gran variedad de estudios independientes	Medio–alto, con múltiples experimentos transcriptómicos	Alto, especialmente en estudios genéticos poblacionales	No aplica	No aplica	No aplica
Actualización y mantenimiento	Actualización continua y mantenimiento activo del portal GDC	Actualización frecuente de nuevas series y estudios	Actualización regular por EMBL-EBI	Actualización periódica de estudios	Actualización continua del catálogo	Actualización continua del catálogo	Actualización continua del agregador
Cumplimiento ético y regulatorio	Muy alto (lineamientos NCI/NIH, datos anonimizados	Adecuado, con políticas NCBI para datos sensibles	Adecuado, alineado a políticas EMBL-EBI	Muy alto (gestión estricta de datos sensibles)	Evalúa y cataloga estándares FAIR	Evalúa y cataloga repositorios FAIR	Cumple directrices europeas de datos abiertos

	, control de acceso)						
Adecuación para integración multimodal	Excelente (clínica + ómica + WSI en una misma cohorte)	Limitada (principalmente ómica)	Limitada (principalmente ómica)	Media (genética + fenotipo, sin imágenes)	No aplica	No aplica	No aplica
Rol en este proyecto	Fuente principal de datos multimodales (TCGA-STAD)	Fuente complementaria para contraste o ampliación ómica	Fuente complementaria para contraste o ampliación ómica	No esencial para el flujo principal	Directorio para identificación de repositorios	Directorio para identificación de repositorios	Directorio/agregador de literatura y datos

El análisis comparativo mostró que TCGA, a través del portal GDC, fue el único repositorio que cumplió simultáneamente con: disponer de información clínica estructurada, ofrecer perfiles genómicos procesables como miRNA-seq y RNA-seq, proporcionar imágenes histopatológicas en formato WSI asociadas a los mismos casos, y mantener una trazabilidad robusta mediante identificadores únicos que permiten vincular las distintas modalidades a nivel de paciente. Estas características, junto con la calidad de los metadatos, el volumen de casos y las sólidas garantías éticas, justificaron su selección como fuente principal de datos para el desarrollo del modelo de predicción de supervivencia.

Por su parte, GEO y ArrayExpress fueron considerados como repositorios complementarios, útiles para contrastar o ampliar los perfiles transcriptómicos en estudios específicos, aunque su falta de integración nativa con imágenes histopatológicas y la variabilidad en la anotación clínica limitaron su papel en el flujo principal de trabajo. En cuanto a dbGaP, si bien ofrece un volumen importante de estudios genéticos con fenotipos asociados y un fuerte componente regulatorio, su enfoque en acceso controlado y la ausencia sistemática de imágenes histopatológicas lo hicieron menos adecuado para los objetivos de integración multimodal.

Finalmente, los directorios FAIRsharing, re3data y OpenAIRE desempeñaron un rol de apoyo metodológico, permitiendo verificar estándares de calidad, localizar repositorios adicionales y asegurar alineación con principios FAIR y políticas de ciencia abierta, más que servir como fuentes directas de datos para el modelado.

5.3. Comprensión de los datos

Datos clínicos

La cohorte clínica incluyó 422 pacientes y 35 variables, cuya estructura y completitud fueron evaluadas para determinar su utilidad en los modelos de supervivencia. En esta etapa se verificó la consistencia de los identificadores (*Case ID* y *Submitter ID*), la disponibilidad de

variables de desenlace (*Vital Status*, *Days to Death*) y la calidad general de la información demográfica, diagnóstica y tumoral.

El análisis inicial mostró que la mayoría de las variables descriptivas fundamentales se encuentran completas, mientras que variables relacionadas con progresión, tratamientos o desenlaces detallados presentan altos niveles de ausencia. Esto permitió anticipar qué campos serían utilizables de manera directa, cuáles requerirían imputación y cuáles no aportarían información suficiente para los modelos.

La Tabla VIII resume las variables, su significado y el porcentaje de datos faltantes, destacando su relevancia para el análisis posterior.

Tabla VIII. Variables clínicas, descripción y porcentaje de datos faltantes

Variable	Descripción	No-Null	% Faltante
Case ID	Identificador único del paciente	422	0%
Submitter ID	Centro que aportó el caso	422	0%
Project ID	Código del proyecto (TCGA-STAD)	422	0%
Gender	Sexo del paciente	422	0%
Age at Index	Edad al diagnóstico	417	1.18%
Country	País reportado	393	6.87%
Vital Status	Estado vital	422	0%
Days to Death	Tiempo hasta la muerte	166	60.66%
Ethnicity	Etnicidad	422	0%
Race	Raza	422	0%
Days to Birth	Días desde el nacimiento al diagnóstico	413	2.13%
Cause of Death	Causa de muerte	166	60.66%
Primary Diagnosis	Diagnóstico principal	422	0%
Tissue Origin	Sitio anatómico	422	0%
AJCC Stage	Estadio clínico	326	22.75%
Year of Diagnosis	Año del diagnóstico	345	18.20%
ICD-10 Code	Código CIE-10	346	18.00%
Tumor Grade	Grado tumoral	343	18.72%
Synchronous Malignancy	Malignidad sincrónica	346	18.00%
Classification of Tumor	Tipo histológico	422	0%
Residual Disease	Enfermedad residual	324	23.22%
Days to Diagnosis	Tiempo al diagnóstico	409	3.08%
Treatment Types	Tipo de tratamiento	385	8.77%
Treatment Intents	Intención terapéutica	258	38.86%
Treatment Outcomes	Resultado del tratamiento	130	69.19%
Therapeutic Agents	Agentes terapéuticos	131	68.96%
Relative with Cancer History	Antecedente familiar	348	17.54%
Relationship Primary Diagnosis	Relación del familiar	348	17.54%
Follow-up Timepoint Category	Categoría de seguimiento	421	0.24%
Disease Response	Respuesta clínica	325	23.22%

Days to Follow-up	Tiempo hasta el seguimiento	416	1.42%
Progression or Recurrence	Progresión/recurrencia	38	91.00%
Progression Type	Tipo de progresión	38	91.00%
Anatomic Site of Progression	Sitio anatómico	38	91.00%
Days to Progression	Tiempo a progresión	25	94.08%

En resumen:

- Variables completas (0%), como *Gender*, *Race*, *Primary Diagnosis* y *Tissue Origin*, ofrecen una base sólida para caracterizar la cohorte.
- Variables críticas con baja ausencia (<5%), como *Age at Index* y *Days to Follow-up*, resultan directamente utilizables y apropiadas para modelos de supervivencia.
- Variables con faltantes moderados (10–30%), especialmente *AJCC Stage*, *Tumor Grade* y *Residual Disease*, requieren imputación o una selección cuidadosa.
- Variables con ausencia severa (>60%), como *Treatment Outcomes* o *Progression Type*, no aportan información confiable y no se consideran en el modelado.

En conjunto, esta etapa permitió identificar qué información clínica era suficientemente robusta para integrarse en los modelos y qué transformaciones serían necesarias en la fase de preparación de datos.

Datos genéticos

La cohorte de datos genómicos corresponde a perfiles de expresión de miRNAs generados a partir de secuenciación de pequeño RNA (small RNA-seq) disponibles en TCGA-STAD. Cada archivo aporta mediciones normalizadas en *reads per million miRNA mapped (RPM)*, una unidad comúnmente utilizada para comparar expresión entre muestras al corregir por la profundidad de secuenciación.
Inicialmente, la estructura del archivo miRNA-seq se extrajo en formato vertical (modelo largo), donde cada fila correspondía a un par (*miRNA*, *Case ID*). Para facilitar su análisis, la matriz se transformó a formato ancho (matriz de expresión), generando un DataFrame con:

- Filas: 436 casos
- Columnas: 1.881 miRNAs
- Valores: niveles de expresión normalizados (RPM)

Tras la depuración y al enlazar estos datos con la cohorte clínica mediante el identificador único *Case ID*, se obtuvo un total de 402 pacientes con información genómica válida. Esto indica que la mayoría de los pacientes con datos clínicos útiles también cuentan con expresión miRNA-seq completa, permitiendo avanzar hacia análisis multivariados consistentes.

La matriz de expresión miRNA-seq presentó un nivel de completitud muy alto, sin valores faltantes, dado que el pipeline de TCGA asigna valores continuos incluso para expresiones muy bajas; aun así, se observaron valores cercanos a cero, marcadas diferencias de escala entre miRNAs altamente expresados y miRNAs raros, y una dispersión amplia típica de datos de secuenciación. La estructura final incluyó cerca de 1882 columnas: una correspondiente al Case ID y 1881 a miRNAs únicos con nomenclatura estandarizada, tales como *hsa-let-7a-1*, *hsa-let-7b*, *hsa-miR-21-5p*, *hsa-miR-200c-3p*, *hsa-miR-451a*, *hsa-miR-125b-1* y otros miembros de familias ampliamente estudiadas en cáncer. Esta dimensionalidad elevada implica correlación entre múltiples miRNAs y resalta la necesidad posterior de aplicar métodos de selección o reducción de características para evitar sobreajuste. Durante la revisión se comprobó la coherencia de las anotaciones, la ausencia de duplicados y la distribución asimétrica esperada de la expresión, sin encontrarse inconsistencias en los identificadores ni problemas en la descarga o lectura inicial de los datos.

Imágenes histopatológicas

Los datos histopatológicos consistieron en 422 casos con Whole Slide Images (WSI) teñidas con hematoxilina y eosina (H&E) provenientes de TCGA-STAD, almacenadas en formato .svs, un estándar de la patología digital que permite múltiples niveles de resolución. Durante la etapa de comprensión se verificaron los metadatos técnicos asociados a cada lámina, incluyendo dimensiones, niveles disponibles, tipo de escáner y resolución efectiva, confirmando su adecuación para análisis computacional a nivel de parches. Asimismo, se revisó la consistencia de los nombres de archivo y su correspondencia con los identificadores del repositorio, garantizando que cada imagen estuviera asociada a un único caso clínico.

En paralelo, se evaluaron características globales del conjunto, observándose una marcada variabilidad entre láminas en términos de tamaño, contraste, condiciones de tinción y presencia de artefactos. Entre los artefactos más frecuentes se identificaron regiones sin tejido, zonas borrosas, áreas sobreexpuestas, fragmentos pequeños o marginales, y diferencias de intensidad relacionadas con el proceso de escaneo. Estas exploraciones preliminares permitieron anticipar la necesidad de aplicar procedimientos posteriores de control de calidad, normalización cromática y segmentación tisular para garantizar la comparabilidad entre casos.

Finalmente, se verificó la trazabilidad entre las imágenes y las demás modalidades de información mediante el Case ID obteniéndose 402 imágenes, confirmando que las WSIs contaban con identificadores consistentes con los utilizados en los datos clínicos y genómicos. Esta correspondencia aseguró que, una vez completada la fase de preparación, sería técnicamente viable integrar la información histopatológica con las matrices clínicas y transcriptómicas, permitiendo el análisis multimodal requerido para los modelos de supervivencia. Los recuentos exactos de WSIs emparejadas con datos clínicos y genéticos

se determinarán en la etapa de resultados, una vez consolidados los filtros de calidad y los criterios de inclusión.

5.4. Preparación de los datos

5.4.1. Preparación de datos clínicos

Análisis de completitud en el conjunto de datos clínico

El análisis de completitud de las variables clínicas presentado en la Fig. 8 evidenció porcentajes elevados de datos faltantes en un conjunto específico de características. En particular, las variables *Days to Progression* (94%), *Anatomic Site of Progression* (91%), *Progression Type* (91%), *Progression or Recurrence*(91%), *Treatment Outcomes* (69%), *Therapeutic Agents* (69%), *Days to Death* (61%), *Cause of Death* (61%), *Residual Disease* (39%) y *Disease Response* (23%) presentaron ausencias significativas. Siguiendo el criterio metodológico definido en la sección 4.4.1. se eliminaron del conjunto aquellas con más del 25% de valores faltantes, lo cual incluyó a todas las variables mencionadas hasta *Residual Disease* inclusive. La variable *Days to Death* fue la única excepción, dado que se mantuvo en el análisis por ser indispensable en el cálculo de la supervivencia del paciente. Esta depuración permitió reducir el sesgo y la incertidumbre asociada a la imputación, garantizando que los análisis posteriores se realizaran sobre un conjunto de datos clínicos más robusto y confiable.

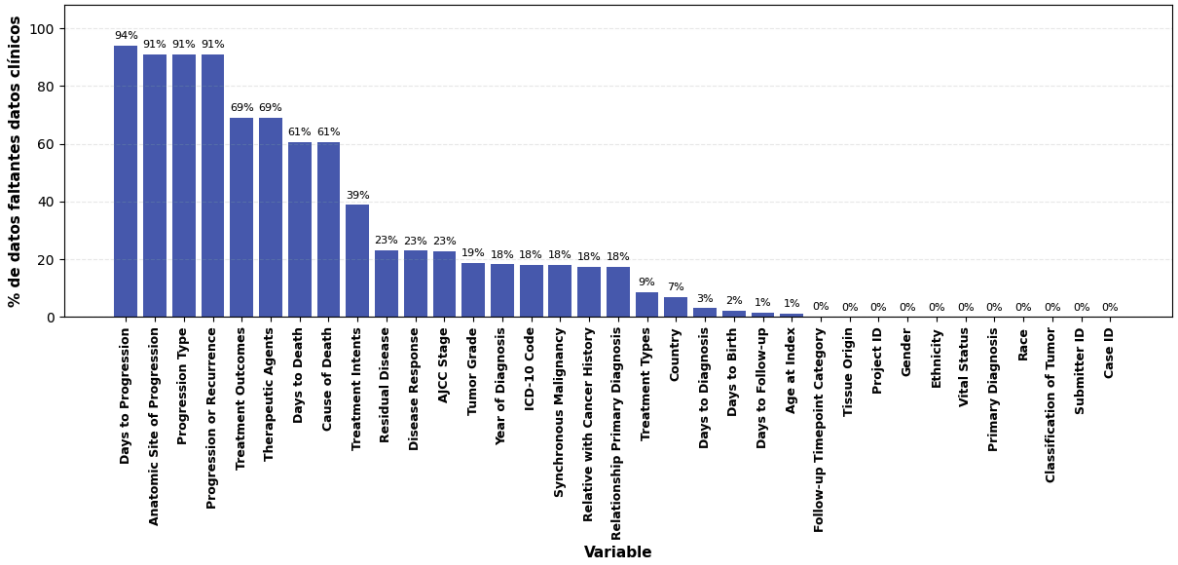


Fig. 8. Análisis de completitud de conjunto de datos clínico TCGA-STAD

Análisis estructural de patrones de ausencia

Como paso inicial para el análisis estructural de patrones de ausencia, se construyó una matriz binaria de ausencia a partir del conjunto de datos clínicos depurado. En esta representación, el valor 1 indica la presencia de un dato faltante y el valor 0 representa un dato observado, lo que facilita el análisis estadístico de patrones de ausencia entre variables. La matriz resultante presentó dimensiones de 415 filas por 24 columnas, correspondientes al número total de pacientes incluidos en el estudio y a las variables clínicas consideradas tras la eliminación de aquellas con más del 25% de valores ausentes.

Esta codificación binaria permitió implementar las métricas estadísticas descritas en la sección 4.4.1. del marco metodológico, sirviendo como base para la evaluación de asociaciones bivariadas entre variables con patrones de ausencia potencialmente dependientes. El análisis subsiguiente se orientó a identificar relaciones estructurales en la falta de datos, con el fin de determinar si existían variables cuya ausencia estuviera asociada de manera significativa a la ausencia en otras, lo que proporcionaría información relevante para el diseño de la estrategia de imputación.

Análisis de correlación entre patrones de ausencia mediante coeficiente ϕ

Para explorar dependencias entre los patrones de ausencia, se construyó una matriz $\phi - \phi$ a partir de los indicadores binarios de faltantes (1 = ausente, 0 = observado). Dado que la correlación de Pearson entre variables binarias es equivalente al coeficiente ϕ (phi) del cruce 2×2, el mapa de calor de la Fig. 9 representa las correlaciones ϕ entre pares de variables. Se observaron bloques coherentes de co-ausencia, por ejemplo, entre *Year of Diagnosis*, *ICD-10 Code* y *Synchronous Malignancy*, lo cual sugiere que algunos patrones de ausencia de información tienden a ocurrir de manera conjunta.

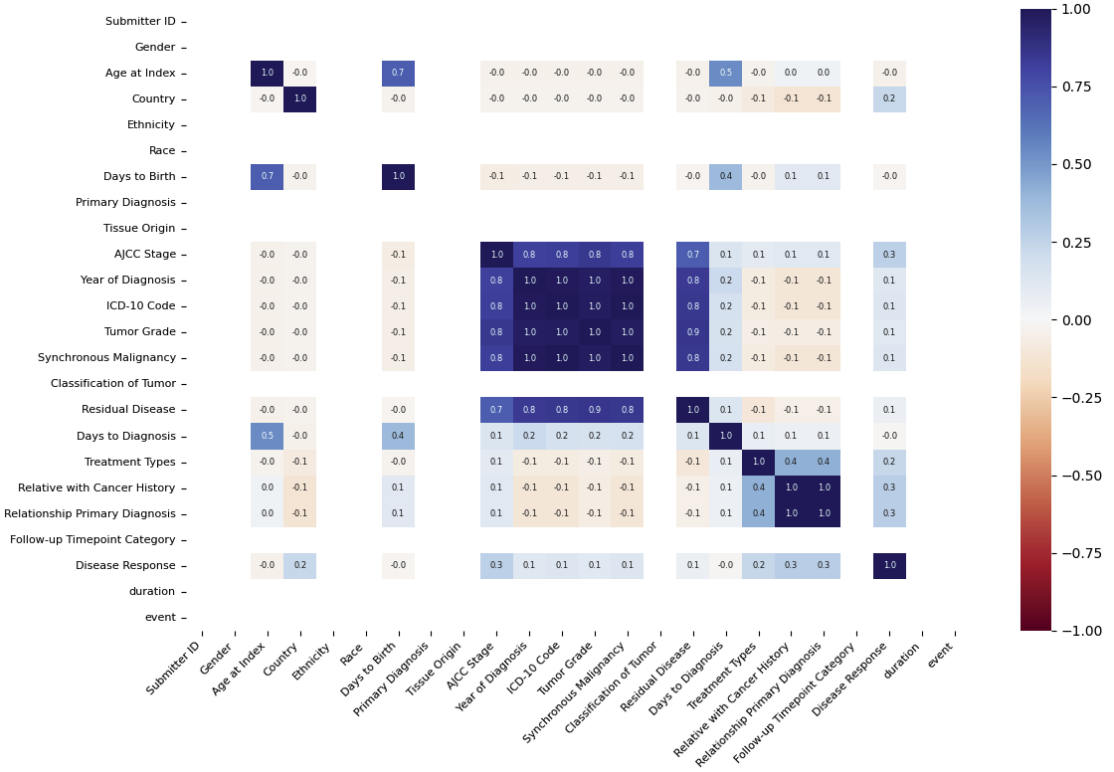


Fig. 9. Matriz ϕ de correlación entre patrones de valores faltantes (datos clínicos TCGA-STAD)

Análisis de asociaciones bivariadas entre patrones de ausencia mediante V de Cramér

A partir de la matriz binaria de valores faltantes se realizó el análisis de asociaciones bivariadas, empleando el coeficiente V de Cramér junto con pruebas de significancia (χ^2 de Pearson o prueba exacta de Fisher, según correspondiera). Los resultados mostraron la existencia de asociaciones muy fuertes entre los patrones de ausencia de varias variables clínicas.

En particular, se identificaron asociaciones perfectas ($V = 1.00; p < 1 \times 10^{-90}$) entre *ICD-10 Code* y *Synchronous Malignancy*, así como entre *Relative with Cancer History* y *Relationship Primary Diagnosis*, lo que indica que estas variables presentan patrones de ausencia idénticos en el conjunto de datos. De manera similar, se observaron asociaciones casi perfectas entre *Year of Diagnosis*, *ICD-10 Code* y *Synchronous Malignancy* ($V \approx 0.99; p < 1 \times 10^{-89}$), lo que sugiere que la información faltante en estas variables sigue una estructura completamente dependiente.

Otras asociaciones destacadas se encontraron entre *Tumor Grade* y *Residual Disease* ($V = 0.85;\ p < 1 \times 10^{-67}$), así como entre *ICD-10 Code*, *Synchronous Malignancy* y *Residual Disease* ($V = 0.85; p < 1 \times 10^{-66}$), lo cual evidencia que la ausencia de datos en variables

clínicas relacionadas con la estadificación y características del tumor tiende a presentarse de manera conjunta. Finalmente, las variables *AJCC Stage*, *Tumor Grade*, *Year of Diagnosis* y *ICD-10 Code* mostraron asociaciones altas (V entre 0.81 y 0.84; $p < 1 \times 10^{-62}$), reforzando la hipótesis de que existen dependencias estructurales en el registro de estas características.

La Tabla IX presenta el resumen de las 15 asociaciones más fuertes identificadas, ordenadas según el valor del V de Cramér. Estos hallazgos permiten concluir que los valores faltantes en varias variables no son independientes, lo cual tiene implicaciones directas sobre el mecanismo de ausencia y justifica el uso de técnicas robustas de imputación, como MICE, en etapas posteriores del análisis.

Tabla IX. Principales asociaciones entre variables clínicas según el estadístico V de Cramér y su significancia estadística

Variable 1	Variable 2	V de Cramér	p-value	Prueba
ICD-10 Code	Synchronous Malignancy	1.000	2.99×10^{-92}	χ^2
Relative with Cancer History	Relationship Primary Diagnosis	1.000	2.99×10^{-92}	χ^2
Year of Diagnosis	ICD-10 Code	0.992	9.40×10^{-91}	χ^2
Year of Diagnosis	Synchronous Malignancy	0.992	9.40×10^{-91}	χ^2
ICD-10 Code	Tumor Grade	0.975	7.04×10^{-88}	χ^2
Tumor Grade	Synchronous Malignancy	0.975	7.04×10^{-88}	χ^2
Year of Diagnosis	Tumor Grade	0.967	1.99×10^{-86}	χ^2
Tumor Grade	Residual Disease	0.853	1.25×10^{-67}	χ^2
ICD-10 Code	Residual Disease	0.846	1.16×10^{-66}	χ^2
Synchronous Malignancy	Residual Disease	0.846	1.16×10^{-66}	χ^2
AJCC Stage	Tumor Grade	0.843	3.08×10^{-66}	χ^2
Year of Diagnosis	Residual Disease	0.838	1.94×10^{-65}	χ^2
AJCC Stage	ICD-10 Code	0.822	6.11×10^{-63}	χ^2
AJCC Stage	Synchronous Malignancy	0.822	6.11×10^{-63}	χ^2
AJCC Stage	Year of Diagnosis	0.814	8.89×10^{-62}	χ^2

Análisis de la fuerza de asociación mediante odds ratio (OR)

Con el objetivo de complementar el análisis de coausencias realizado mediante el coeficiente ϕ y el estadístico V de Cramér, se evaluó la fuerza de asociación entre los patrones de ausencia de las variables clínicas utilizando el estadístico *odds ratio* (OR). Esta métrica estima la probabilidad relativa de que la ausencia de una variable ocurra simultáneamente con la ausencia de otra, en comparación con la probabilidad de que

ambas ocurran de manera independiente. Asimismo, se calcularon intervalos de confianza al 95% y se aplicaron pruebas χ^2 de Pearson o exactas de Fisher, según la distribución esperada de frecuencias.

Los resultados obtenidos revelaron asociaciones extremadamente fuertes entre diversos pares de variables. Destaca la relación entre *Relative with Cancer History* y *Relationship Primary Diagnosis*, que presentó un OR ≈ 100 695.0 (IC95%: 1981.8 – 5.1×10^6; $p < 0.001$), lo que indica que la probabilidad de que ambas variables estén ausentes simultáneamente es más de 100 000 veces mayor que la esperada por azar. De forma similar, la coausencia entre *ICD-10 Code* y *Synchronous Malignancy* fue altamente significativa (OR ≈ 99 615.0; IC95%: 1960.3 – 5.06×10^6; $p < 0.001$), evidenciando una fuerte dependencia estructural entre ambas.

También se observaron asociaciones notables entre *Year of Diagnosis* y *ICD-10 Code* (OR ≈ 33 108.3; IC95%: 1335.2 – 8.20×10^5; $p < 0.001$), así como entre *Tumor Grade* y *Synchronous Malignancy* (OR ≈ 14 106.4; IC95%: 720.7 – 2.76×10^7; $p < 0.001$). La relación entre *Tumor Grade* y *Residual Disease* (OR ≈ 1184.0; IC95%: 156.4 – 8963.4; $p < 0.001$) y entre *ICD-10 Code* y *Residual Disease* (OR ≈ 2071.9; IC95%: 124.2 – 3455.4; $p < 0.001$) refuerzan la evidencia de que la ausencia de información en variables asociadas al diagnóstico y características del tumor no ocurre de manera independiente.

Finalmente, aunque con una magnitud considerablemente menor, la relación entre *Relative with Cancer History* y *Disease Response* (OR ≈ 4.33; IC95%: 2.53 – 7.40; $p < 0.001$) sugiere la existencia de asociaciones moderadas en variables clínicas relacionadas con el seguimiento del paciente. En conjunto, estos resultados confirman que los valores faltantes en el conjunto clínico de TCGA-STAD presentan dependencias estructurales significativas, lo cual respalda la decisión metodológica de aplicar técnicas de imputación multivariante en etapas posteriores del análisis.

La Tabla X resume las principales asociaciones identificadas, junto con sus intervalos de confianza y valores de significancia estadística.

Tabla X. Principales asociaciones entre patrones de valores faltantes expresadas como odds ratio, con intervalos de confianza del 95% y significancia estadística

Variable 1	**Variable 2**	***Odds Ratio***	**IC 95% Inferior**	**IC 95% Superior**	***p*-valor**	**Prueba**
Relative with Cancer History	Relationship Primary Diagnosis	100 695.00	1 981.80	5.12×10^6	2.99×10^{-92}	χ^2
ICD-10 Code	Synchronous Malignancy	99 615.00	1 960.38	5.06×10^6	2.99×10^{-92}	χ^2
Year of Diagnosis	ICD-10 Code	33 108.33	1 335.23	8.21×10^5	9.40×10^{-91}	χ^2

Year of Diagnosis	Synchronous Malignancy	33 108.33	1 335.23	8.21×10^{5}	9.40×10^{-91}	χ^2
ICD-10 Code	Tumor Grade	14 106.43	720.77	2.76×10^{5}	7.04×10^{-88}	χ^2
Tumor Grade	Synchronous Malignancy	14 106.43	720.77	2.76×10^{5}	7.04×10^{-88}	χ^2
Year of Diagnosis	Tumor Grade	8 136.00	834.30	7.93×10^{4}	1.99×10^{-86}	χ^2
Tumor Grade	Residual Disease	1 184.00	156.40	8.96×10^{3}	1.25×10^{-67}	χ^2
ICD-10 Code	Residual Disease	2 071.89	124.23	3.46×10^{4}	1.16×10^{-66}	χ^2
Synchronous Malignancy	Residual Disease	2 071.89	124.23	3.46×10^{4}	1.16×10^{-66}	χ^2
AJCC Stage	Tumor Grade	584.00	133.52	2.55×10^{3}	3.08×10^{-66}	χ^2
Year of Diagnosis	Residual Disease	1 047.27	138.89	7.90×10^{3}	1.94×10^{-65}	χ^2
AJCC Stage	ICD-10 Code	486.96	112.20	2.11×10^{3}	6.11×10^{-63}	χ^2
AJCC Stage	Synchronous Malignancy	486.96	112.20	2.11×10^{3}	6.11×10^{-63}	χ^2
AJCC Stage	Year of Diagnosis	323.62	94.53	1.11×10^{3}	8.89×10^{-62}	χ^2
AJCC Stage	Residual Disease	46.75	24.38	89.64	1.52×10^{-46}	χ^2
Treatment Types	Relative with Cancer History	15.67	7.25	33.88	1.22×10^{-17}	χ^2
Treatment Types	Relationship Primary Diagnosis	15.67	7.25	33.88	1.22×10^{-17}	χ^2
Relative with Cancer History	Disease Response	4.33	2.53	7.41	1.99×10^{-8}	χ^2
Relationship Primary Diagnosis	Disease Response	4.33	2.53	7.41	1.99×10^{-8}	χ^2

Análisis visual de patrones de ausencia

El análisis visual de los valores faltantes permitió examinar la estructura y posibles dependencias entre las ausencias del conjunto clínico. Para ello, se emplearon herramientas gráficas como *Missingno* y *UpSetPlot*, que facilitan la identificación de relaciones de co-ocurrencia entre las variables y los patrones sistemáticos de ausencia. Estas representaciones complementan los análisis estadísticos previos (ϕ, V de Cramér y Odds Ratio) y ayudan a determinar si las ausencias pueden considerarse aleatorias o si, por el contrario, responden a mecanismos dependientes de otras variables.

La Fig. 10 presenta la matriz de presencia/ausencia, que ilustra de manera global la distribución de los valores observados y faltantes. Se aprecia una concentración marcada de ausencias en variables vinculadas con el diagnóstico y la caracterización tumoral, como *ICD-10 Code*, *Synchronous Malignancy*, *Tumor Grade*, *Residual Disease* y *AJCC Stage*, mientras que las variables demográficas (*Gender*, *Age at Index*, *Race*, *Ethnicity*) presentan alta completitud. Estos patrones indican que la pérdida de información no es uniforme entre las observaciones, sino que se concentra en dimensiones clínicas específicas, reflejando posiblemente diferencias en los protocolos de registro o en la disponibilidad de datos hospitalarios.

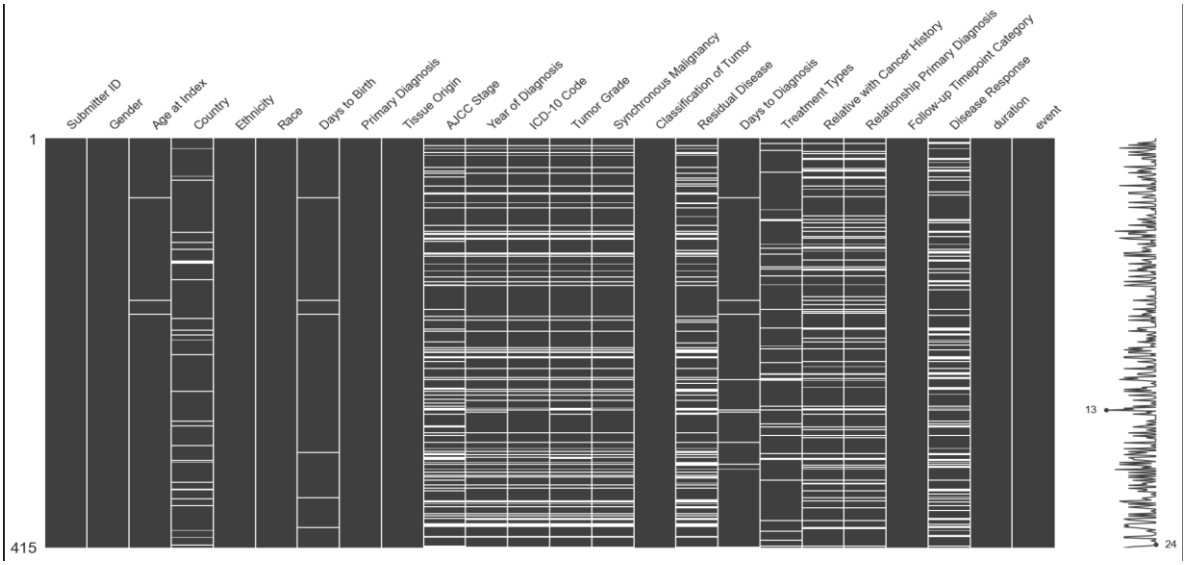


Fig. 10. Matriz de valores faltantes del conjunto de datos clínico TCGA-STAD

La Fig. 11 presenta el dendrograma que agrupa las variables en función de la similitud de sus patrones de ausencia. Se identifican tres conglomerados principales:

- G**rupo diagnóstico** (*Year of Diagnosis*, *ICD-10 Code*, *Synchronous Malignancy*, *Tumor Grade*, *Residual Disease*, *AJCC Stage*).
- G**rupo de historia clínica y tratamiento** (*Relative with Cancer History*, *Relationship Primary Diagnosis*, *Treatment Types*).
- G**rupo de variables completas o casi completas** (*Gender*, *Age at Index*, *Race*, *Country*).

La agrupación evidencia dependencias estructuradas entre las ausencias, coherentes con los resultados del heatmap de la Fig. 9.

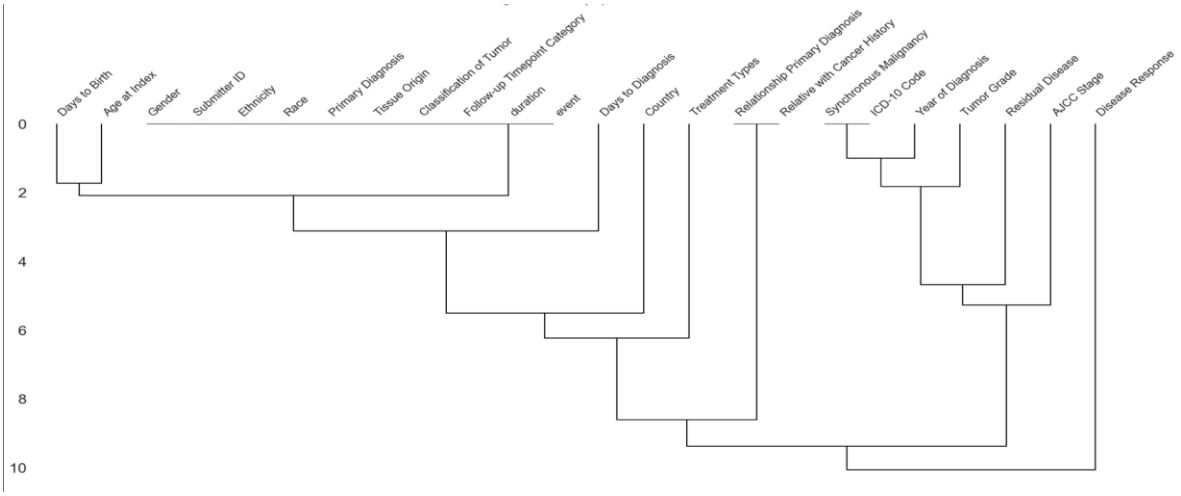


Fig. 11. Dendrograma jerárquico por similitud de patrones de faltantes

En la Fig. 12, el diagrama UpSet cuantifica las combinaciones de variables que presentan valores faltantes simultáneamente. Las intersecciones más frecuentes corresponden a *Disease Response*, *Residual Disease*, *AJCC Stage*, *Tumor Grade*, *ICD-10 Code* y *Synchronous Malignancy*, que concentran la mayor proporción de ausencias. Este tipo de visualización resalta que las pérdidas de información no ocurren de forma aleatoria, sino que tienden a agruparse en subconjuntos de variables clínicas relacionadas.

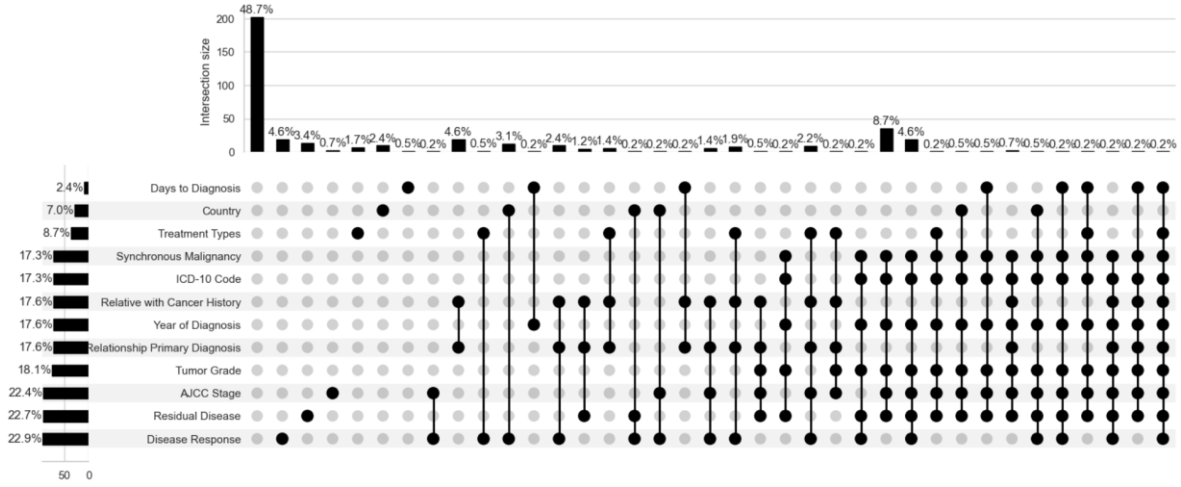


Fig. 12. Diagrama UpSet de combinaciones de variables con faltantes

En conjunto, los resultados obtenidos a partir de las visualizaciones evidencian que los valores faltantes en el conjunto clínico TCGA-STAD no siguen un patrón completamente aleatorio (MCAR).
Las ausencias muestran una estructura definida y correlacionada entre variables del

dominio diagnóstico, lo que sugiere la presencia de un mecanismo MAR (Missing At Random).
Por este motivo, se procedió a aplicar la prueba MCAR de Little para confirmar estadísticamente esta hipótesis y orientar la estrategia de imputación multivariante utilizada en los análisis posteriores.

Aplicación de prueba MCAR de Little

Con el fin de verificar si los valores faltantes del conjunto clínico TCGA-STAD seguían un patrón completamente aleatorio (MCAR, *Missing Completely At Random*), se aplicó la prueba de Little (Little's MCAR Test), una extensión multivariante de la prueba χ^2 que contrasta la hipótesis nula de aleatoriedad total en los datos ausentes.

Previo a la prueba, las variables categóricas fueron transformadas a códigos numéricos, se identificaron los patrones de ausencia y se agruparon las observaciones con estructuras de faltantes similares. A partir de las medias y covarianzas del conjunto completo, se calculó el estadístico χ^2 acumulado considerando el tamaño de cada grupo y la distancia de sus medias con respecto a la media global.

Los resultados obtenidos fueron los siguientes:

$$\chi^2 = 70.6140, gl = 24, p = 1.76 \times 10^{-6}$$

El valor p extremadamente bajo ($p < 0.001$) llevó a rechazar la hipótesis nula de que los valores faltantes son completamente aleatorios. Por tanto, se concluye que los datos no son MCAR, sino que probablemente siguen un mecanismo MAR (Missing At Random), en el que la probabilidad de ausencia depende de otras variables observadas. Este hallazgo es consistente con los resultados previos del análisis gráfico y estadístico que evidenciaron una co-ocurrencia estructurada de ausencias entre variables clínicas relacionadas con el diagnóstico y el tratamiento.

En consecuencia, se descartó el uso de métodos de imputación simples (como la media o la moda) y se optó por aplicar imputación múltiple por ecuaciones encadenadas (MICE), la cual preserva las relaciones multivariadas entre variables y proporciona estimaciones más robustas frente a mecanismos MAR.

Imputación de valores faltantes mediante el método MICE

Una vez identificados los patrones de ausencia y determinado que los datos no son completamente al azar (MCAR), se procedió a la imputación de valores faltantes mediante el método Multiple Imputation by Chained Equations (MICE). Este enfoque se basa en la idea de generar múltiples estimaciones posibles de los valores ausentes a partir de modelos iterativos que aprovechan las relaciones existentes entre las variables.

En cada iteración, MICE ajusta un modelo de regresión para cada variable con valores faltantes, utilizando como predictores las demás variables del conjunto de datos. Posteriormente, los valores imputados se actualizan de forma encadenada hasta lograr la convergencia. Esta estrategia permite conservar la **estructura multivariada y las interdependencias naturales de los datos**, evitando los sesgos introducidos por métodos más simples como la imputación por la media o la eliminación de casos incompletos.

En el presente estudio, el método MICE se aplicó tanto a variables numéricas como categóricas, utilizando codificación ordinal para estas últimas y un máximo de diez iteraciones por cadena. Este procedimiento permitió obtener un conjunto de datos imputado **completo y coherente**, apto para el desarrollo posterior de los modelos de predicción de supervivencia.

Con el fin de evaluar el impacto de la imputación y verificar que el proceso no alterara de manera significativa las distribuciones originales, se elaboraron matrices de gráficos comparativos entre las versiones del conjunto de datos antes y después del proceso de imputación.

En la Fig. 13 se comparan las distribuciones de las variables numéricas con valores faltantes antes y después de la imputación. Se observa que las curvas de densidad de variables como *Age at Index*, *Days to Birth* y *Days to Diagnosis* presentan una superposición casi total entre ambas versiones del conjunto de datos. Esto indica que la imputación con MICE preservó la forma y tendencia original de las distribuciones, sin generar desplazamientos evidentes en la media ni incrementos artificiales en la varianza.

De igual modo, la variable *Year of Diagnosis* mantiene su concentración principal alrededor del mismo rango temporal, mostrando que el proceso de imputación fue coherente con la estructura temporal observada previamente en los datos completos.

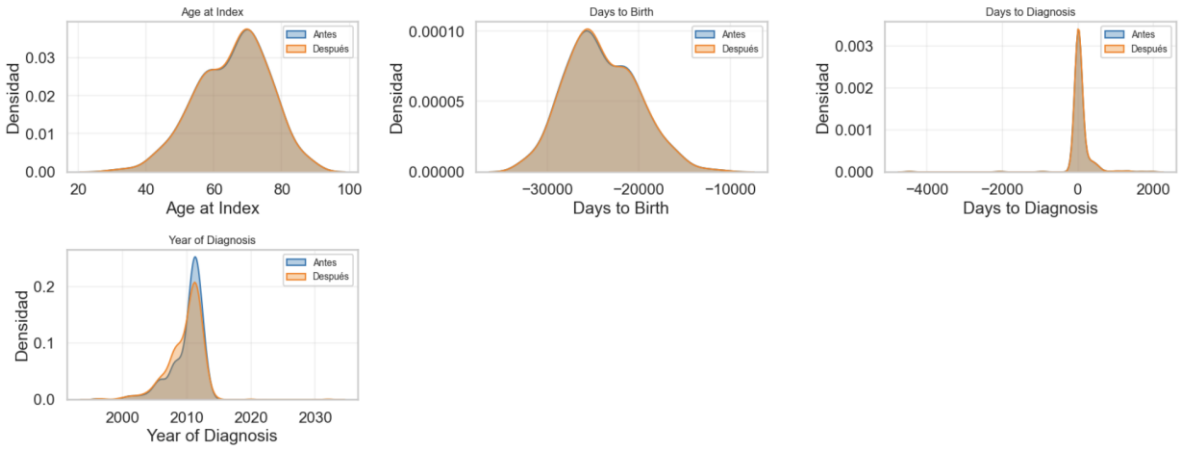


Fig. 13. Comparación de las distribuciones de variables numéricas antes y después de la imputación MICE

En la Fig. 14 se presentan las distribuciones porcentuales de las variables categóricas antes y después de la imputación. Se observa que las proporciones entre categorías permanecen estables para variables relevantes como *AJCC Stage*, *Tumor Grade*, *Disease Response*, *Residual Disease* y *Synchronous Malignancy*. Por ejemplo, la imputación no generó redistribuciones extremas en los estadios clínicos ni en los tipos de respuesta a la enfermedad, sino que repartió los valores faltantes siguiendo patrones plausibles en función de las relaciones identificadas entre variables.

En el caso de *Treatment Types* y *Country*, se observan ligeras variaciones en categorías con baja frecuencia, pero estas permanecen dentro de un margen razonable, lo que sugiere que la imputación mantuvo la coherencia estructural del conjunto de datos.

En conjunto, los resultados de las comparaciones gráficas confirman que el procedimiento de imputación no distorsionó de manera significativa las distribuciones originales de las variables ni generó sesgos aparentes. Esto valida la adecuación del método MICE para este conjunto de datos, garantizando la conservación de la información y la consistencia interna entre variables. Además, al contrastar los resultados con las pruebas previas, se refuerza la conclusión de que los datos presentan mecanismos de ausencia no completamente aleatorios (MAR o MNAR), razón por la cual la imputación múltiple resultó el enfoque más apropiado para su tratamiento antes del modelado posterior de supervivencia.

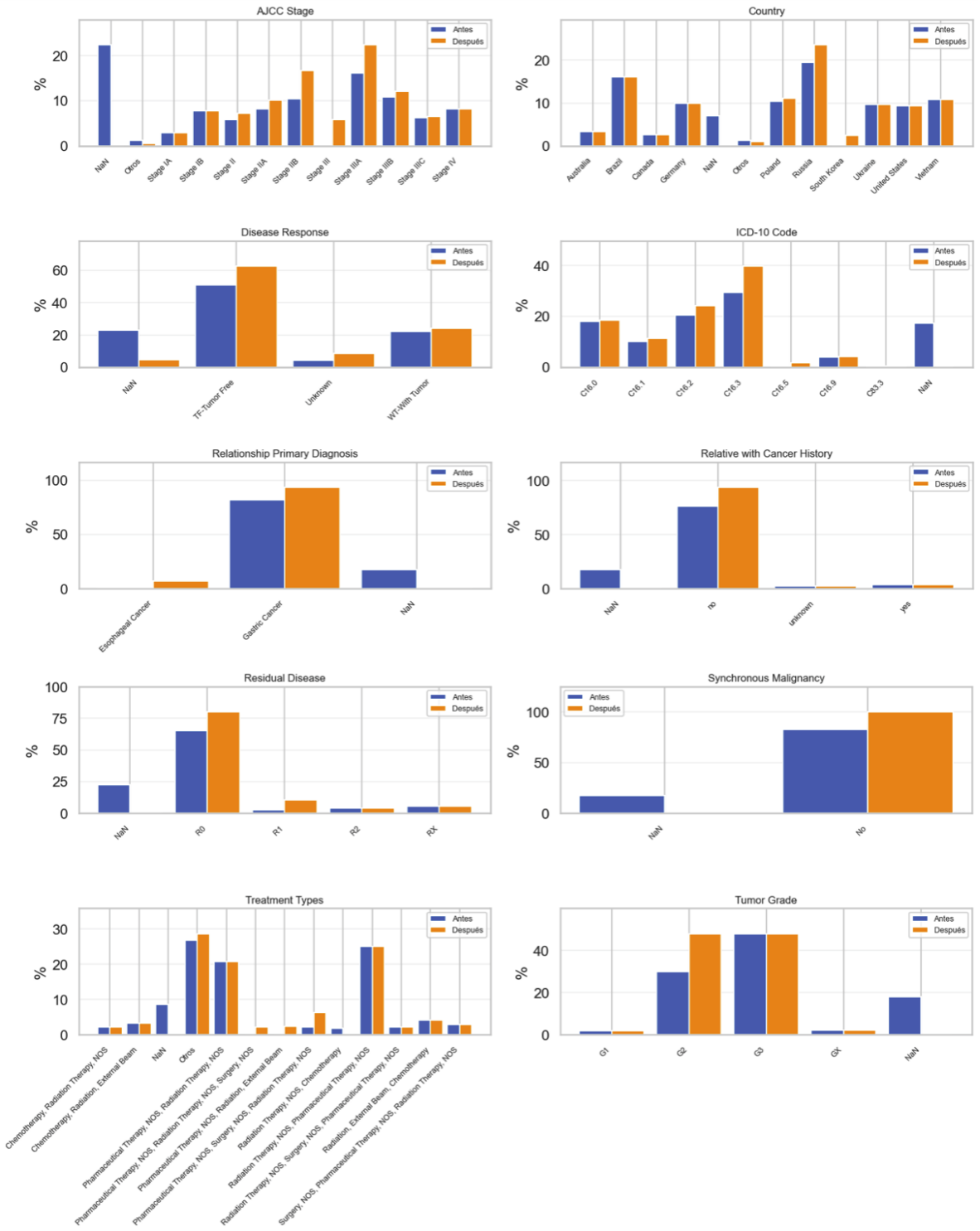


Fig. 14. Comparación de las distribuciones de variables categóricas antes y después de la imputación MICE

5.4.2. Preparación de datos genéticos

El conjunto miRNA-seq incluyó 436 pacientes y 1881 miRNAs, sin valores faltantes y sin duplicados, lo que confirma que los pasos previos del pipeline de TCGA como filtrado de lecturas, eliminación de adaptadores y secuencias de baja calidad, mapeo al genoma de referencia y detección de duplicados, produjeron una matriz depurada y estructuralmente consistente. La normalización aplicada por TCGA también se mantuvo intacta, asegurando comparabilidad entre muestras.

Los datos mostraron el patrón típico de la expresión miRNA-seq: 67,2 % de valores iguales a 0 RPM, con un promedio de 1264 miRNAs no expresados por paciente. Este comportamiento refleja la biología del transcriptoma de miRNA y no deficiencias técnicas. Asimismo, el tamaño de biblioteca fue exactamente 1.000.000 RPM en todas las muestras, confirmando una normalización homogénea y sin distorsiones derivadas de diferencias en profundidad de secuenciación.

En términos de variabilidad, 484 miRNAs presentaron varianza muy baja (<0.001), incluyendo transcritos como *hsa-miR-921*, *hsa-miR-4307* o *hsa-miR-4309*. Adicionalmente, la mayoría de los miRNAs mostró valores con entre 7 y 10 decimales, debido a la escala RPM, sin representar inconsistencias. En conjunto, estos resultados indican que los datos genéticos poseen alta calidad, homogeneidad y estabilidad numérica, y que no fue necesario excluir miRNAs ni pacientes en esta etapa, cuyo objetivo era únicamente preparar el conjunto de datos antes de su uso en análisis posteriores.

5.4.3. Preparación de imágenes histopatológicas

Para la preparación del conjunto de imágenes histopatológicas se realizó un filtrado previo de aquellas que no cumplían con las condiciones piramidales de magnificación como un nivel compatible a 20X, como resultado de este ejercicio se obtuvieron 386 WSIs, a partir de las cuales se generó automáticamente un conjunto de etiquetas de calidad mediante la inspección visual asistida por el entorno gráfico de HistoQC y los comentarios generados manualmente. Como resultado, 258 láminas (66.8%) fueron clasificadas como *Ok*, 128 (33.2%) como *Review* y ninguna como *Bad*, de modo que únicamente las láminas *Ok* y *Review* fueron empleadas para las etapas posteriores de extracción de parches.

Con el fin de homogenizar la variabilidad cromática entre láminas, Se seleccionó el caso TCGA-VQ-A8PE visualizado en la Fig. 15 como referencia debido a su tinción representativa basada en métricas como *H_mean*, *E_mean*, *varianza* y *porcentaje de tejido*. Dicha lámina se utilizó como referencia para aplicar la normalización de color mediante el método de Macenko, lo que produjo un conjunto final de imágenes normalizadas cromáticamente, con tinciones H&E más consistentes entre pacientes y sin alteraciones estructurales visibles.

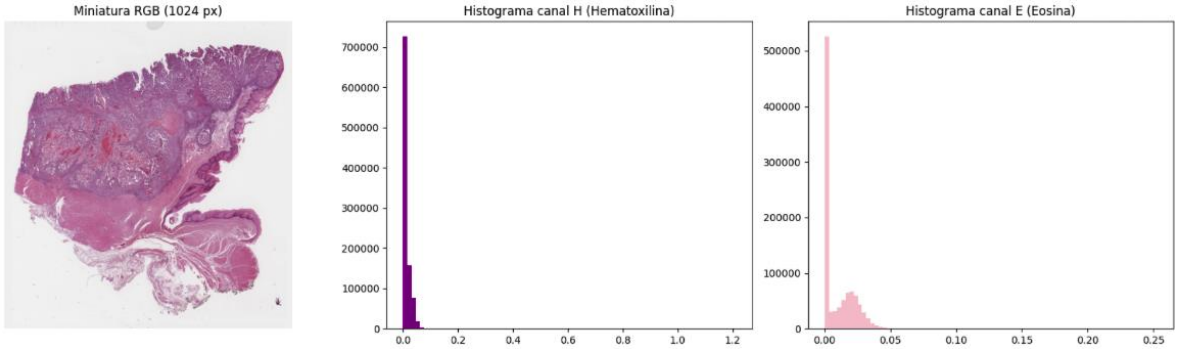


Fig. 15. WSI de referencia para normalizador Macenko

Finalmente, se procedió a la extracción de parches. Para cada WSI válida, se generaron parches de 1024×1024 píxeles, con un traslape del 25%, todos ellos ya normalizados en color y etiquetados según la categoría de calidad original (*Ok* o *Review*). Posteriormente, se aplicó un filtro morfológico que retuvo únicamente aquellos parches con ≥50% de contenido tisular, descartando regiones predominantemente blancas o carentes de estructura. Este proceso produjo el conjunto de parches histopatológicos que se utilizaría en las etapas posteriores de extracción y selección de características.

5.5. Selección de características

5.5.1. Selección de características clínicas

Siguiendo el criterio metodológico definido en la sección 4.5.1., los resultados del análisis univariado se presentan en la Tabla XI, donde se observa que las variables *Disease Response*, *Residual Disease*, *AJCC Stage* y *Follow-up Timepoint Category* mostraron asociaciones estadísticamente significativas con la supervivencia ($p < 0.05$).

Entre ellas, *Disease Response* ($HR = 1.97$, $p \approx 1.94x10^{-17}$, $C - index = 0.66$) y *Residual Disease* ($HR = 1.91$, $p \approx 1.28x10^{-14}$, $C - index = 0.61$) las más fuertemente asociadas con el riesgo de mortalidad, indicando que una peor respuesta al tratamiento o presencia de enfermedad residual incrementan significativamente el riesgo de muerte. Por su parte, *AJCC Stage* ($HR = 1.17$, $p \approx 2.91x10^{-6}$) confirmó su importancia pronóstica, en concordancia con la literatura clínica que destaca el estadio tumoral como uno de los factores determinantes de la supervivencia en cáncer gástrico.

Tabla XI. Resultados del modelo de regresión de Cox univariado para las variables clínicas

Variable	Coef	Exp(coef)	*p*	*p* < 0.05	C-Index	% NA	Error
Disease Response	0.676033	1.966062	1.946419e-17	True	0.659416	0.0	None
Residual Disease	0.645687	1.907298	1.283892e-14	True	0.605398	0.0	None

AJCC Stage	0.158416	1.171653	2.914508e-06	True	0.616068	0.0	None
Follow-up Timepoint Category	0.152195	1.164387	2.248953e-02	True	0.530748	0.0	None
Submitter ID	-0.001256	0.998745	5.004858e-02	False	0.520915	0.0	None
Days to Birth	-0.000041	0.999959	5.497882e-02	False	0.534395	0.0	None
Age at Index	0.014912	1.015023	5.573480e-02	False	0.535445	0.0	None
Ethnicity	0.203083	1.225174	1.092149e-01	False	0.524349	0.0	None
Country	0.031792	1.032303	1.192261e-01	False	0.524150	0.0	None
Tumor Grade	0.215854	1.240921	1.199286e-01	False	0.554274	0.0	None
Days to Diagnosis	-0.000196	0.999804	2.689652e-01	False	0.512118	0.0	None
Tissue Origin	0.014722	1.014830	3.364334e-01	False	0.528066	0.0	None
Primary Diagnosis	-0.013819	0.986276	4.324597e-01	False	0.528478	0.0	None
Classification of Tumor	0.068422	1.070817	4.456916e-01	False	0.508301	0.0	None
Race	0.034497	1.035099	4.592144e-01	False	0.512316	0.0	None
Treatment Types	0.003379	1.003385	5.417180e-01	False	0.521355	0.0	None
Year of Diagnosis	0.010760	1.010818	6.481579e-01	False	0.473750	0.0	None
ICD-10 Code	0.014559	1.014666	8.074896e-01	False	0.511905	0.0	None
Gender	0.038340	1.039085	8.155295e-01	False	0.497730	0.0	None
Relative with Cancer History	-0.004357	0.995653	9.786265e-01	False	0.506541	0.0	None
Relationship Primary Diagnosis	-0.006274	0.993746	9.861998e-01	False	0.498808	0.0	None
Synchronous Malignancy	NaN	NaN	NaN	False	NaN	0.0	Convergence halted due to matrix inversion problema.

Aunque variables como *Age at Index* y *Tumor Grade* no alcanzaron significancia estadística ($p > 0.05$), se mantuvieron en la selección final debido a su relevancia clínica y su uso frecuente en modelos pronósticos según lo presentado en la sección 3.1.2.3..

En conjunto, las variables seleccionadas para el análisis multivariado posterior fueron: *Disease Response*, *Residual Disease*, *AJCC Stage*, *Tumor Grade*, *Age at Index*, *Gender*, *Tissue Origin*, *Classification of Tumor* y *Treatment Types*.

5.5.2. Selección de características genéticas

Para la selección de características genéticas se ajustó un modelo de regresión de Cox multivariado empleando las variables clínicas seleccionadas en la sección 5.5.1.. El ajuste se realizó mediante el estimador de Breslow con penalización de 0.1, considerando 415 observaciones **y** 166 eventos de mortalidad.

El modelo alcanzó un índice de concordancia (*C-index*) de 0.81, lo que indica una alta capacidad discriminativa para diferenciar entre pacientes con mayor o menor riesgo de mortalidad.
Asimismo, la prueba de razón de verosimilitud (*likelihood ratio test*) resultó altamente significativa ($\chi^2 = 161.17, p < 0.001$), evidenciando que el conjunto de covariables contribuye significativamente a la explicación del riesgo de muerte.

Las variables con mayor impacto pronóstico fueron *Disease Response*, *Residual Disease* y *AJCC Stage*, todas con coeficientes positivos y valores de p estadísticamente significativos.
En particular, los pacientes con enfermedad residual RX (HR = 4.31, p < 0.001) o R2 (HR = 2.94, p = 0.01) mostraron un riesgo de mortalidad considerablemente superior en comparación con aquellos con resección completa (R0). De igual modo, una respuesta al tratamiento tipo "With Tumor" (HR = 2.08, p < 0.001) duplicó el riesgo de mortalidad frente al grupo "Tumor-Free".

Los estadios avanzados (*AJCC III–IV*) también incrementaron significativamente el riesgo (HR = 1.67, p = 0.02), mientras que los estadios tempranos (*IA y IB*) mostraron un efecto protector.
Por otra parte, la variable Age at Index (HR = 1.02, p = 0.01) evidenció que el riesgo de mortalidad aumenta ligeramente con la edad. En contraste, variables como *Gender* y *Tissue Origin* no resultaron significativas en el modelo global, aunque algunos subtipos anatómicos presentaron tendencias relevantes a nivel descriptivo.

La Tabla XII resume las covariables más influyentes del modelo multivariado, destacando aquellas con significancia estadística o relevancia clínica.

Tabla XII. Covariables más relevantes en el modelo multivariado de Cox

Variable	Coeficiente	HR (exp(coef))	IC95% HR inferior	IC95% HR superior	p-valor	Significancia
Disease Response – WT (With Tumor)	0.73	2.08	1.43	3.03	<0.001	***

Disease Response – Unknown	0.55	1.74	0.99	3.03	0.05	*
Residual Disease – R2	1.08	2.94	1.28	6.75	0.01	**
Residual Disease – RX	1.46	4.31	2.04	9.08	<0.001	***
AJCC Stage – III / IV	0.51	1.67	1.15	3.12	0.02	**
Tumor Grade – GX	0.78	2.17	0.70	6.78	0.10	—
Age at Index	0.02	1.02	1.01	1.04	0.01	**

Nota: se presentan únicamente las covariables con significancia estadística ($p < 0.05$) o relevancia clínica destacada.

Estratificación por grupos de riesgo y estadio

A partir de los coeficientes del modelo se calculó un índice de riesgo individual (risk score) para cada paciente, mediante la combinación ponderada de las covariables clínicas. La distribución del índice permitió estratificar a los pacientes en dos grupos pronósticos:

- **Bajo riesgo (n = 208)**: pacientes con menor probabilidad de mortalidad.
- **Alto riesgo (n = 207)**: pacientes con mayor riesgo de mortalidad esperada según el modelo.

La curva de Kaplan–Meier presentada en la Fig. 16 mostró una separación clara y estadísticamente significativa entre ambos grupos (log-rank p = 7.18e–33). El grupo de alto riesgo experimentó una disminución rápida de la probabilidad de supervivencia en los primeros tres años, alcanzando una mediana cercana a los 2 años, mientras que el grupo de bajo riesgo mantuvo una probabilidad de supervivencia superior al 60% incluso después de 5 años.

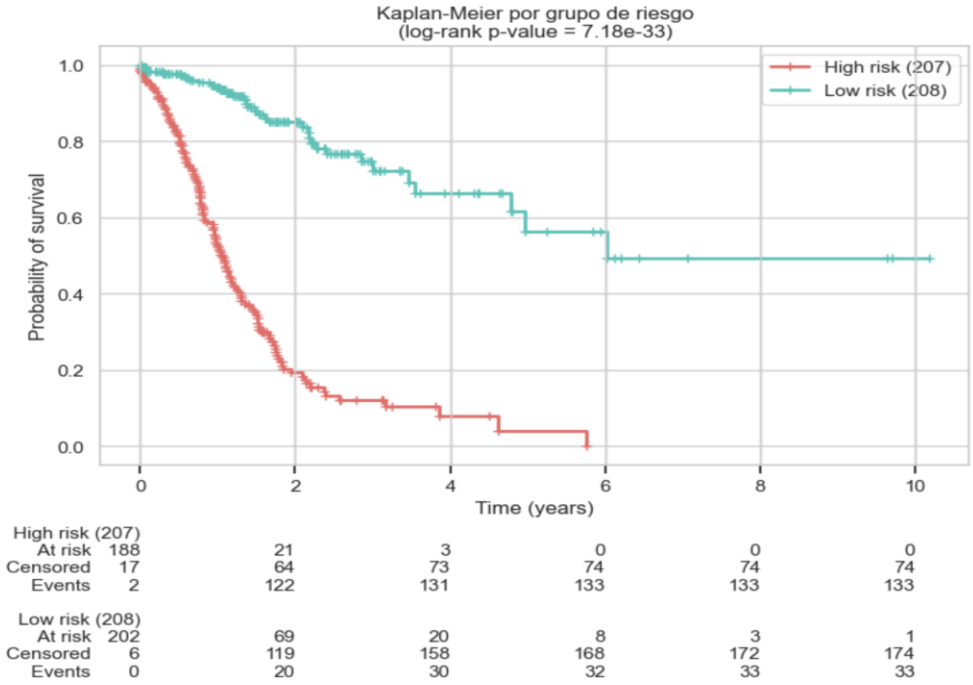


Fig. 16. Curva Kaplan–Meier de supervivencia estratificada por grupos de riesgo obtenidos del modelo Cox multivariado

Estos resultados confirman la validez del modelo multivariado para establecer una asociación robusta de los pacientes según su riesgo clínico, la cual se tomó como base para la asociación posterior con los perfiles genéticos (miRNA**)** en los análisis de expresión diferencial mediante el cálculo de *Fold Change* y *p-value*.

Paralelamente, se decidió replicó este ejercicio tomando como grupos de comparación los estadios tempranos del cáncer (I y II) vs estadios avanzados (III y IV) debido a su importancia clínica según la literatura presentada en la sección 3.1.2.3..

Análisis de expresión diferencial y filtrado final con modelos Cox univariados.

Los resultados obtenidos del filtrado realizado mediante análisis de expresión génica diferencial siguiendo los lineamientos metodológicos propuestos en la sección 4.5.2. se presentan en la Tabla XIII para la comparación por grupo de riesgo y en la Tabla XIV para la comparación entre estadios tempranos y avanzados.

Tabla XIII. Micro RNA obtenidos del filtrado mediante expresión génica diferencial comparando los grupos low risk y high risk

mi-RNA ID	MATURE 5â€™ arm	MATURE 3â€™ arm	Mean LowR	Mean HighR	FC	t-test (p)
hsa-mir-548d-2	hsa-miR-548d-5p	hsa-miR-548d-3p	21609	48961	2.27	0.00
hsa-mir-548e	hsa-miR-548e-3p	hsa-miR-548e-5p	15397	52462	3.41	0.00
hsa-mir-4640	hsa-miR-4640-5p	hsa-miR-4640-3p	0	0	8.23	0.00
hsa-mir-3615	hsa-miR-3615		496	5544	11.19	0.01
hsa-mir-6732	hsa-miR-6732-5p	hsa-miR-6732-3p	1970	8614	4.37	0.01
hsa-mir-374b	hsa-miR-374b-5p	hsa-miR-374b-3p	2018	15492	7.68	0.01
hsa-mir-671	hsa-miR-671-5p	hsa-miR-671-3p	28215	56532	2.00	0.01
hsa-mir-548d-1	hsa-miR-548d-5p	hsa-miR-548d-3p	4348	12330	2.84	0.01
hsa-mir-544b	hsa-miR-544b		11565	25123	2.17	0.01
hsa-mir-548ag-1	hsa-miR-548ag		679	9933	14.63	0.01
hsa-mir-3119-2	hsa-miR-3119		0	0	4.55	0.01
hsa-mir-4644	hsa-miR-4644		6835	15852	2.32	0.01
hsa-mir-3147	hsa-miR-3147		0	0	7.04	0.02
hsa-mir-3936	hsa-miR-3936		1551	10983	7.08	0.02
hsa-mir-5685	hsa-miR-5685		504	3696	7.33	0.02
hsa-mir-5587	hsa-miR-5587-5p	hsa-miR-5587-3p	804	4722	5.87	0.02
hsa-mir-4674	hsa-miR-4674		0	0	3.24	0.02
hsa-mir-5002	hsa-miR-5002-5p	hsa-miR-5002-3p	4212	11165	2.65	0.02
hsa-mir-5197	hsa-miR-5197-5p	hsa-miR-5197-3p	19079	39507	2.07	0.03
hsa-mir-3134	hsa-miR-3134		0	0	3.12	0.03
hsa-mir-611	hsa-miR-611		1281	7915	6.18	0.03
hsa-mir-516b-2	hsa-miR-516b-5p	hsa-miR-516b-3p	8853	24990	2.82	0.04
hsa-mir-1537	hsa-miR-1537-3p	hsa-miR-1537-5p	497	2942	5.91	0.04
hsa-mir-4537	hsa-miR-4537		0	0	10.50	0.04
hsa-mir-558	hsa-miR-558		7545	16946	2.25	0.04
hsa-mir-6124	hsa-miR-6124		2405	6589	2.74	0.04
hsa-mir-1185-1	hsa-miR-1185-5p	hsa-miR-1185-1-3p	545	3460	6.35	0.04
hsa-mir-3683	hsa-miR-3683		8287	16817	2.03	0.04
hsa-mir-7843	hsa-miR-7843-5p	hsa-miR-7843-3p	3592	10456	2.91	0.05
hsa-mir-6743	hsa-miR-6743-5p	hsa-miR-6743-3p	4935	16247	3.29	0.05
hsa-mir-4686	hsa-miR-4686		3415	8857	2.59	0.05
hsa-mir-7111	hsa-miR-7111-5p	hsa-miR-7111-3p	2410	7644	3.17	0.05
hsa-mir-1273h	hsa-miR-1273h-5p	hsa-miR-1273h-3p	10047	22685	2.26	0.05
hsa-mir-4532			640	4195	6.56	0.05
hsa-mir-4783	hsa-miR-4783-5p	hsa-miR-4783-3p	5473	12882	2.35	0.05
hsa-mir-320d-2	hsa-miR-320d		1090	4176	3.83	0.05

Tabla XIV. Micro RNA obtenidos del análisis de expresión génica comparando los grupos por estadios I y II vs III y IV

mi-RNA ID	Mean Stage III and IV	Mean Stage I and II	FC	t-test(p)
hsa-mir-1323	3113,55	16554	5,32	0,03
hsa-mir-3143	1040,08	5707,15789	5,49	0,03
hsa-mir-3145	2323,94	11575,0426	4,98	0,01
hsa-mir-3679	2348,40	8453,35971	3,60	0,02
hsa-mir-4433b	1765,06	6099,28481	3,46	0,05
hsa-mir-498	1813,23	13409,791	7,40	0,03
hsa-mir-5092	5756,70	18263,9922	3,17	0,02
hsa-mir-5094	898,07	4426,12179	4,93	0,04
hsa-mir-515-1	1880,02	11926,6111	6,34	0,05
hsa-mir-518ª-1	4147,09	17123,5079	4,13	0,05
hsa-mir-520e	2721,60	18081,8384	6,64	0,04
hsa-mir-6807	2960,92	10272,1667	3,47	0,04
hsa-mir-6821	243,74	2324,04375	9,53	0,05

Una vez reducida la alta dimensionalidad de los datos genéticos se realizó un último filtrado mediante modelos de cox univariados para cada grupo obteniéndose los siguientes mi-RNA como significativos para la predicción de la supervivencia como se observa en la Tabla XV.

Tabla XV. Micro RNA significativos mediante modelos de Cox univariados

Covariable	Beta	HR	p	Obtenido por
hsa-mir-6732	0.148228	1.159778	0.006885	Grupo de riesgo
hsa-mir-3143	0.163313	1.177405	0.005736	Estadio
hsa-mir-4674	-0.233324	0.791897	0.033645	Grupo de riesgo

Los modelos de Cox aplicados a los miRNAs seleccionados permitieron identificar asociaciones significativas entre niveles de expresión y riesgo de mortalidad. Entre ellos, hsa-miR-6732 mostró un efecto de riesgo elevado (β=0.148, HR=1.159, p=0.0069), indicando que una mayor expresión se relaciona con incremento en la probabilidad de muerte. De manera similar, hsa-miR-3143 presentó un efecto comparable (β=0.163, HR=1.177, p=0.0057), coherente con una mayor presencia en tumores clínicamente más avanzados. En contraste, hsa-miR-4674 evidenció un comportamiento protector (β=−0.233, HR=0.792, p=0.0336), asociado a una reducción del riesgo de mortalidad en pacientes con mayor expresión de este marcador. Estos resultados reflejan patrones diferenciales de riesgo vinculados a la expresión de miRNAs específicos y sustentan su potencial relevancia pronóstica en el contexto del cáncer gástrico.

5.5.3. Extracción y selección de características histopatológicas

En la fase de extracción, a partir de los parches tisulares normalizados se generó una matriz con 38 descriptores histopatológicos por parche, que incluyen métricas de morfometría epitelial (n_islas, med_area, med_circularidad, med_elongacion), arquitectura glandular (gland_count, gland_area_frac), desmoplasia y estroma (desmo_score), necrosis e inflamación perinecrótica (nec_area_frac, infl_perinec_density), mucina (mucin_area_frac, frac_mucina), fracción epitelial (frac_epitelio), así como características de calidad/localización del frente tumoral y textura (borde_sharpness, borde_edge_density, entropy_all, energy_all, attention_score, bright_clip, dark_clip, tile_blur). La matriz resultante presentó prácticamente ausencia de valores faltantes y mantuvo la trazabilidad por paciente y por lámina, lo que permitió consolidar un conjunto consistente de características histológicas computables a gran escala.

Para hacer comparables las láminas entre sí, las características calculadas por parche se agregaron a nivel de paciente mediante estadísticas robustas: media, mediana, percentiles superiores (p90, p95, p99) y máximos. De esta forma, cada descriptor histopatológico quedó representado por varios resúmenes que capturan tanto el comportamiento central de la distribución como la presencia de valores extremos (por ejemplo, desmo_score_median, desmo_score_p90, borde_sharpness_mean, tile_blur_p99, infl_perinec_density_mean), generando un perfil cuantitativo global de cada tumor y su microambiente.

6. MODELADO Y EVALUACIÓN

Continuando con las etapas de modelado y de evaluación según la metodología CRISP-DM, tras completar la etapa de selección de características presentada en las secciones 5.5.1., 5.5.2. y 5.5.3., se implementaron los modelos de supervivencia Regresión Cox, Random Survival Forest y DeepSurv para los conjuntos de datos clínico, clínico + genético miRNA y clínico + genético miRNA + histopatológico donde se compararon dichos modelos y las métricas de evaluación.

Modelos de supervivencia basados en conjunto de datos clínico

Los tres modelos evaluados se optimizaron siguiendo procedimientos específicos a su naturaleza. El Coxnet penalizado utilizó una arquitectura lineal con regularización elástica, cuyo espacio de hiperparámetros fue ajustado mediante búsqueda aleatoria (RandomizedSearchCV) sobre las proporciones de penalización l1_ratio y alpha_min_ratio, lo que permitió controlar el sobreajuste en un conjunto clínico de dimensionalidad moderada. El Random Survival Forest (RSF) empleó una estructura no paramétrica basada en 300 árboles y se ajustó mediante búsqueda aleatoria sobre profundidad máxima, número mínimo de muestras por división y número de estimadores, optimizando su

capacidad para capturar interacciones no lineales entre predictores clínicos. Por su parte, DeepSurv implementó una red neuronal totalmente conectada, regularizada con dropout y early stopping, y optimizada con el algoritmo Adam; su estructura se ajustó mediante un proceso iterativo de validación interna para estabilizar el comportamiento del riesgo continuo.

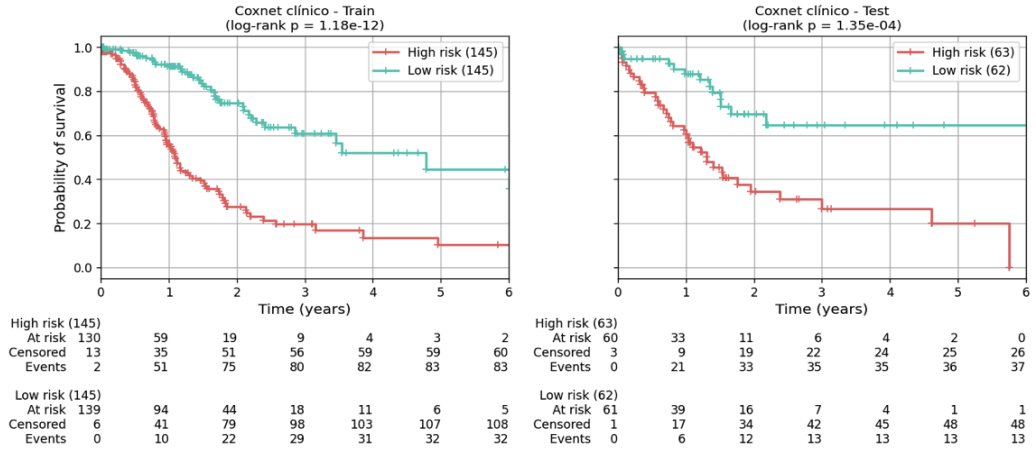
Los valores obtenidos en la Tabla XVI de desempeño muestran que los tres modelos alcanzaron una capacidad discriminativa consistente en el conjunto de prueba, con C-index entre 0.686 y 0.701. El Coxnet penalizado logró la mejor discriminación (0.7008), seguido de DeepSurv (0.6928) y RSF (0.6863). Los Brier Scores, en el rango 0.152–0.158, indican una precisión razonable en la estimación de supervivencia, destacando nuevamente a Coxnet como el modelo más equilibrado en prueba. Todos los enfoques produjeron log-rank p extremadamente significativos, confirmando que las predicciones de riesgo generaron grupos clínicamente distinguibles. En entrenamiento, los modelos mostraron mejor rendimiento con C-index entre 0.731 y 0.772, especialmente el RSF, lo que sugiere cierto grado de sobreajuste inherente a su naturaleza no paramétrica.

Tabla XVI. Desempeño de modelos en conjunto de datos clínico

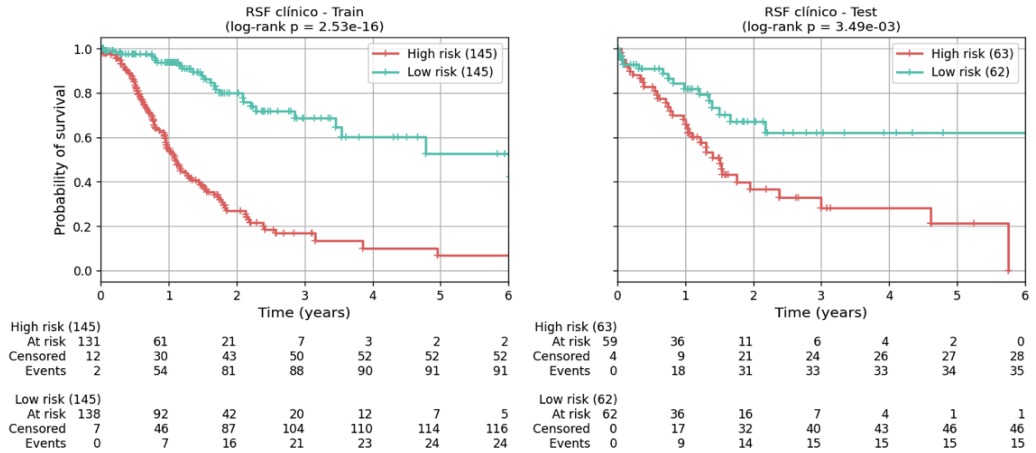
Conjunto	Modelo	C-index	Brier Score
Test	Coxnet Penalizado	0.7008	0.1527
	Random Survival Forest	0.6863	0.1576
	DeepSurv	0.6928	0.1536
Train	Coxnet Penalizado	0.7311	0.1332
	Random Survival Forest	0.7719	0.1249
	DeepSurv	0.7537	0.1254

Las curvas de Kaplan–Meier generadas a partir de las predicciones (Fig. 17) permitieron visualizar la separación entre los grupos de alto y bajo riesgo definidos por cada modelo. Tanto en entrenamiento como en prueba, la divergencia entre las curvas fue amplia y consistente, reflejando patrones clínicos esperables en cáncer gástrico, donde el estadio avanzado, la mala respuesta terapéutica y la mayor agresividad tumoral se asocian a reducciones marcadas en supervivencia. Los log-rank p observados (< 0.001 en todos los casos) respaldan estadísticamente esta separación, evidenciando que cada modelo fue capaz de estratificar correctamente a los pacientes en dos trayectorias de supervivencia clínicamente diferenciables.

A)



B)



C)

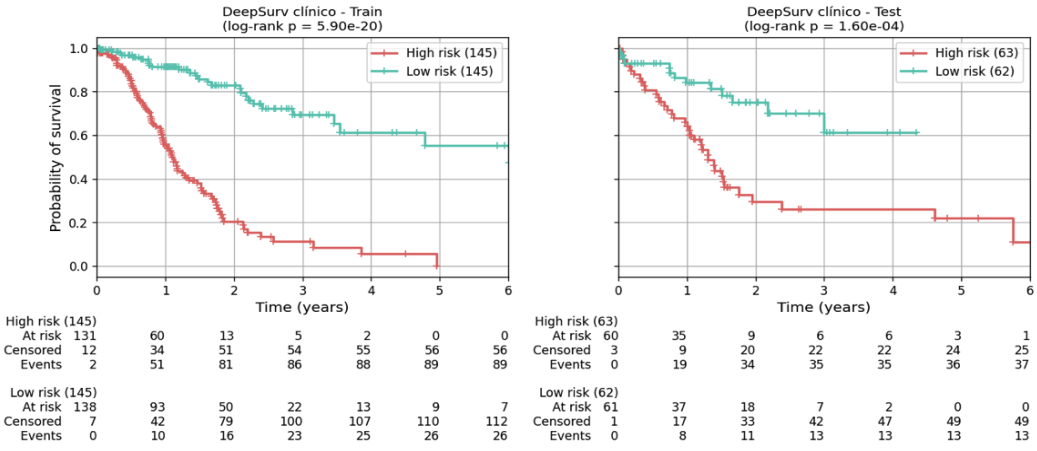


Fig. 17. Curvas de Kaplan–Meier para los modelos implementados en el conjunto de datos clínico

Las curvas de calibración a 1, 3 y 5 años (Fig. 18) mostraron una correspondencia aceptable entre las probabilidades predichas y las observadas. El Coxnet penalizado y el RSF exhibieron la mejor alineación con la diagonal ideal, particularmente en los cuantiles intermedios, lo que indica una adecuada estimación del riesgo. DeepSurv, en contraste, tendió a sobreestimar el riesgo en los cuantiles superiores, un comportamiento coherente con la mayor flexibilidad de su arquitectura y con la menor estabilidad de las redes neuronales en escenarios con proporciones moderadas de censura. No obstante, en conjunto, los tres modelos demostraron una calibración clínicamente aceptable.

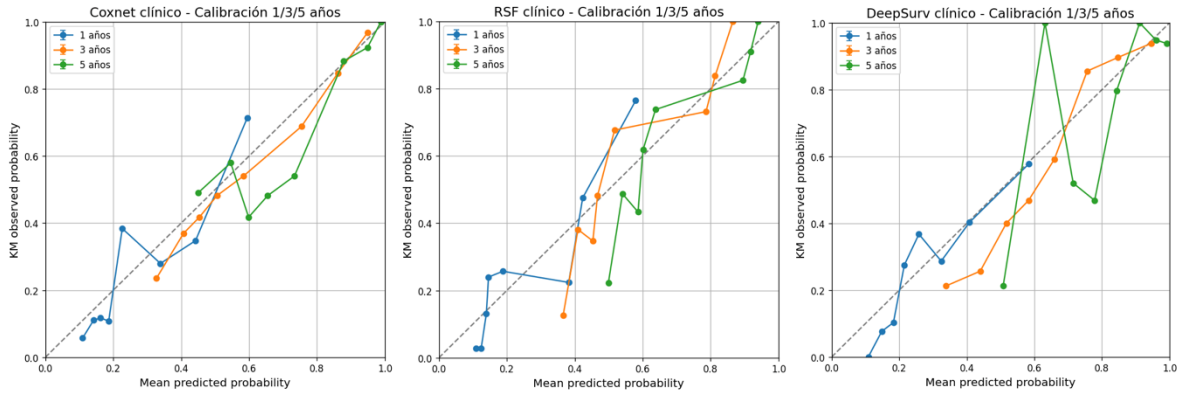


Fig. 18. Curvas de calibración para los modelos implementados en el conjunto de datos clínico

El análisis de las curvas ROC dependientes del tiempo (Fig. 19) mostró que los tres modelos mantuvieron un desempeño similar durante el primer año, con AUC(t) ≈ 0.70–0.73. Sin embargo, al evaluar la evolución temporal, el Coxnet penalizado mantuvo los valores más estables en horizonte de 3 y 5 años (AUC ≈ 0.757–0.763), seguido del RSF (AUC ≈ 0.742–0.745). DeepSurv presentó una tendencia decreciente después del tercer año (de 0.751 a 0.644) (Fig. 20), lo que evidencia una menor estabilidad temporal de la red neuronal frente a modelos más rígidos o regularizados. Este comportamiento sugiere que, en el contexto puramente clínico, las técnicas lineales penalizadas y los métodos ensamblados ofrecen una mejor generalización temporal.

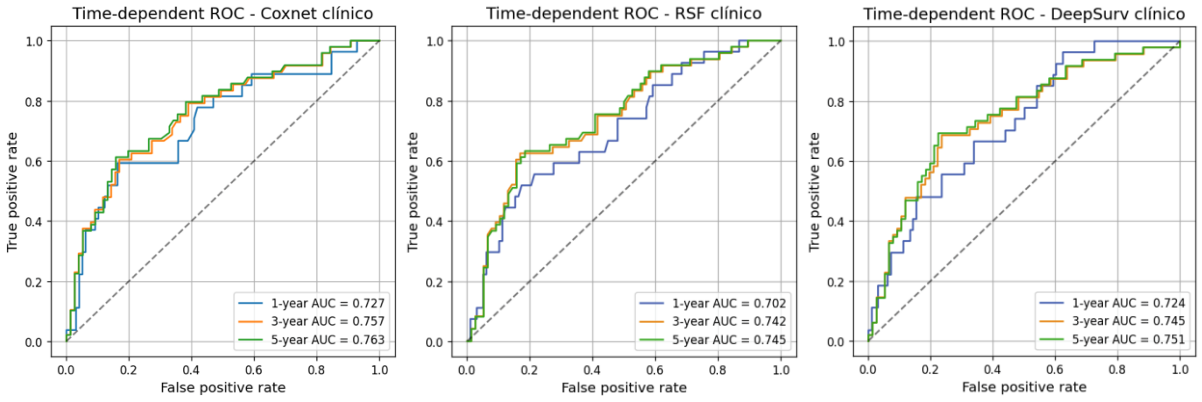


Fig. 19. Curvas ROC dependientes del tiempo para los modelos implementados para el conjunto de datos clínico

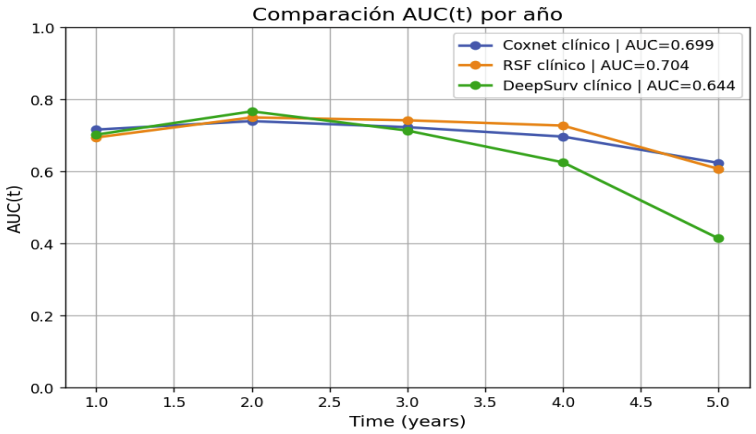


Fig. 20. Comparación de AUC(t) para los modelos implementados para el conjunto de datos clínico

El análisis comparativo de las matrices de confusión para los horizontes de 1, 3 y 5 años (Fig. 21) mostró que Coxnet, Random Survival Forest (RSF) y DeepSurv presentan comportamientos diferenciados en la identificación de pacientes con riesgo de mortalidad en cáncer gástrico. En el horizonte de 1 año, Coxnet obtuvo el menor número de falsos negativos, evidenciando una mayor sensibilidad para detectar eventos tempranos, mientras que DeepSurv y RSF presentaron desempeños similares aunque ligeramente menos sensibles. A los 3 años, tanto Coxnet como DeepSurv mostraron un equilibrio superior entre verdaderos positivos y verdaderos negativos, superando a RSF, que registró más falsos negativos y una menor capacidad discriminativa. Esta tendencia se mantuvo en el horizonte de 5 años, donde Coxnet y DeepSurv conservaron un desempeño prácticamente idéntico, exhibiendo mayor sensibilidad y precisión en comparación con RSF. En conjunto, los resultados indican que Coxnet es el modelo más estable y consistente a través de todos los horizontes, mientras que DeepSurv iguala su rendimiento en proyecciones de mediano plazo gracias a su capacidad para capturar relaciones no lineales en los datos. RSF, aunque aceptable, mostró la mayor tasa de falsos negativos, lo que sugiere que es menos adecuado en contextos clínicos donde la subestimación del riesgo puede comprometer la toma de decisiones.

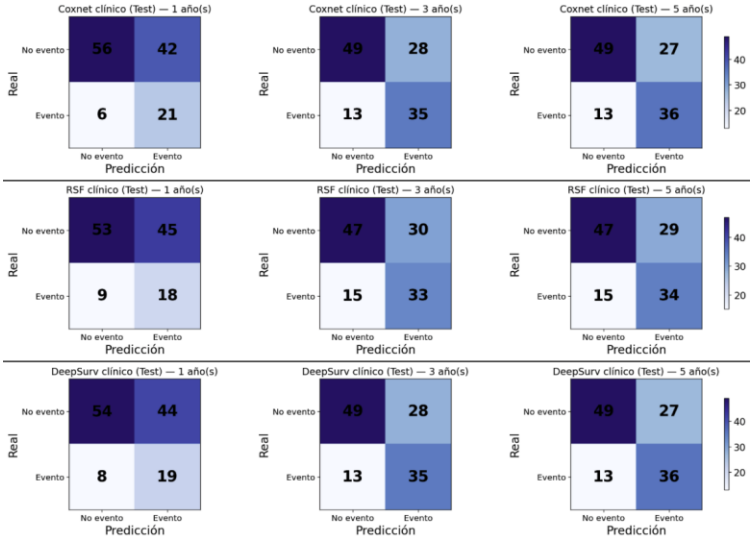


Fig. 21. Matrices de confusión para 1, 3 y 5 años para los modelos implementados para el conjunto de datos clínico

Tomando en conjunto las métricas de discriminación, calibración, estabilidad temporal y significancia estadística, el Coxnet penalizado se posiciona como el modelo clínico más equilibrado y robusto. Aunque el RSF mostró un C-index superior en entrenamiento, su caída en prueba y su comportamiento menos estable en horizontes largos sugieren mayor sensibilidad al sobreajuste. DeepSurv logró un buen desempeño inicial, pero su reducción de AUC(t) después del tercer año limita su utilidad como herramienta longitudinal. Por su consistencia global, su sólida calibración y su adecuado rendimiento en todos los horizontes temporales evaluados, el Coxnet penalizado se considera el mejor modelo clínico para la predicción de supervivencia en esta cohorte TCGA-STAD.

Modelos de supervivencia basados en conjunto de datos clínico y genético miRNA

Al integrar las variables clínicas con los miRNA seleccionados anteriormente, se ajustaron nuevamente los tres modelos de supervivencia utilizando procedimientos de optimización consistentes pero adaptados al aumento de dimensionalidad. El Coxnet penalizado se entrenó con una representación lineal regularizada mediante elastic net, seleccionando automáticamente el nivel de penalización óptimo a través de una búsqueda aleatoria, donde el mejor modelo utilizó l1_ratio = 0.5 y alpha_min_ratio = 0.1. El Random Survival Forest implementó un ensamble de 500 árboles, optimizado mediante random search sobre profundidad máxima (max_depth=4), número de muestras mínimas por hoja y proporción de características (max_features=0.5), buscando capturar interacciones no lineales entre predictores clínicos y moleculares. Por su parte, DeepSurv empleó una red neuronal totalmente conectada con dropout y early stopping, entrenada con el optimizador Adam y validación interna continua, lo que permitió estabilizar el riesgo continuo aun con el incremento en la complejidad del espacio de características.

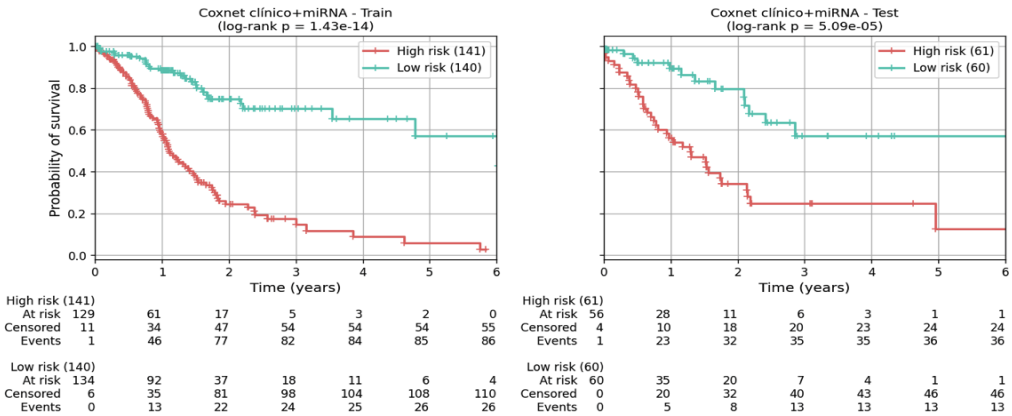
Los resultados cuantitativos (Tabla XVII) muestran una mejora consistente respecto a los modelos exclusivamente clínicos. En el conjunto de prueba, el desempeño fue: C-index entre 0.731 y 0.748, Brier Score entre 0.139 y 0.144, indicando un aumento en capacidad discriminativa atribuible a la inclusión de miRNAs. El RSF obtuvo el mejor C-index en prueba (0.7483), seguido de DeepSurv (0.7430) y Coxnet (0.7315). No obstante, el modelo con mejor Brier Score y mayor equilibrio general fue DeepSurv (0.1394). Los log-rank p del conjunto de prueba fueron altamente significativos en todos los modelos ($p < 1e-05$), confirmando que la combinación clínico+miRNA produjo agrupaciones de riesgo bien diferenciadas. En entrenamiento, el RSF alcanzó nuevamente el mejor C-index (0.7725), mientras que DeepSurv mostró el mejor balance entre discriminación, error de predicción y estabilidad del riesgo.

Tabla XVII. Desempeño de modelos en conjunto de datos clínico y genético (miRNA)

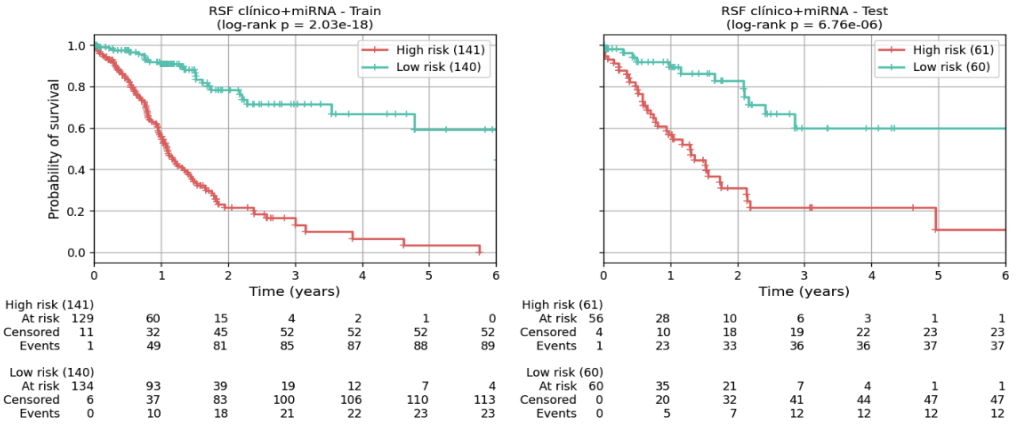
Conjunto	Modelo	C-index	Brier Score
Test	Coxnet Penalizado	0.7315	0.1441
	Random Survival Forest	0.7483	0.1421
	DeepSurv	0.7430	0.1394
Train	Coxnet Penalizado	0.7128	0.1407
	Random Survival Forest	0.7725	0.1217
	DeepSurv	0.7247	0.1331

Las curvas Kaplan–Meier (Fig. 22) mostraron separaciones aún más pronunciadas entre los grupos de alto y bajo riesgo respecto a los modelos clínicos. En ambos subconjuntos (train y test), los tres enfoques generaron curvas bien divergentes, con reducciones marcadas de supervivencia en el grupo de alto riesgo, lo cual respalda el valor pronóstico adicional de los miRNA. Los log-rank p de cada modelo fueron altamente significativos tanto en entrenamiento ($p \approx 10^{-14}–10^{-18}$) como en prueba ($p \approx 10^{-05}–10^{-06}$), evidenciando que la combinación clínico+miRNA mejora la capacidad de estratificación, especialmente en los primeros 3 años, que son críticos en cáncer gástrico avanzado.

A)



B)



C)

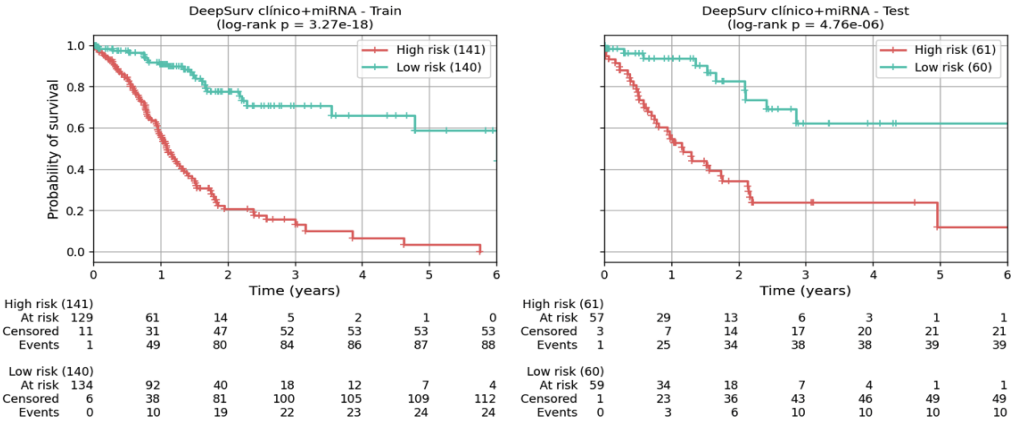


Fig. 22. Curvas de Kaplan–Meier para los modelos implementados en el conjunto de datos clínico y genético (mi-RNA)

Las curvas de calibración (Fig. 23) mostraron un ajuste aceptable en los tres modelos, con tendencia a una mejora en el alineamiento respecto al análisis clínico únicamente. Coxnet y RSF mantuvieron una correspondencia más consistente con la diagonal ideal en horizontes de 3 y 5 años, mientras que DeepSurv mostró una calibración particularmente estable a 1 año, aunque con mayor variabilidad en los cuantiles superiores a 5 años. En conjunto, la calibración confirma que la inclusión de miRNA no introdujo sobreajustes significativos y, por el contrario, contribuyó a una mejor correspondencia entre riesgo predicho y observado.

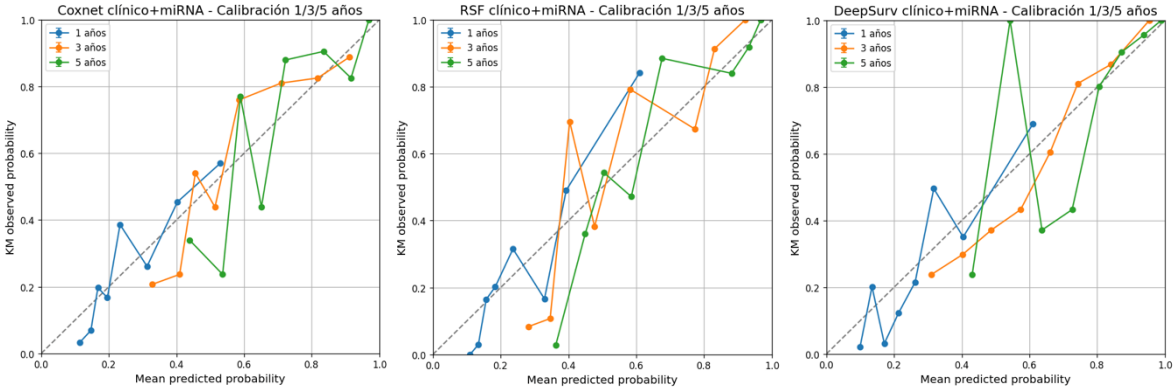


Fig. 23. Curvas de calibración para los modelos implementados en el conjunto de datos clínico y genético (mi-RNA)

El análisis de AUC dependiente del tiempo (Fig. 25 y Fig. 25) mostró que los tres modelos alcanzaron valores superiores a los obtenidos con datos clínicos solos. En el primer año, DeepSurv obtuvo el mejor desempeño (AUC = 0.820), seguido de RSF (0.799) y Coxnet (0.784). A 3 y 5 años, RSF y DeepSurv mantuvieron los valores más altos (0.763–0.767), mientras que Coxnet presentó un AUC ligeramente menor (0.758–0.760). Sin embargo, la comparación global de AUC(t) muestra que DeepSurv exhibió la curva más estable en el tiempo (AUC global ≈ 0.727), seguido de RSF (0.718) y Coxnet (0.669). Esto sugiere que la combinación clínico+miRNA es especialmente beneficiosa para modelos con capacidad no lineal o jerárquica como RSF y DeepSurv.

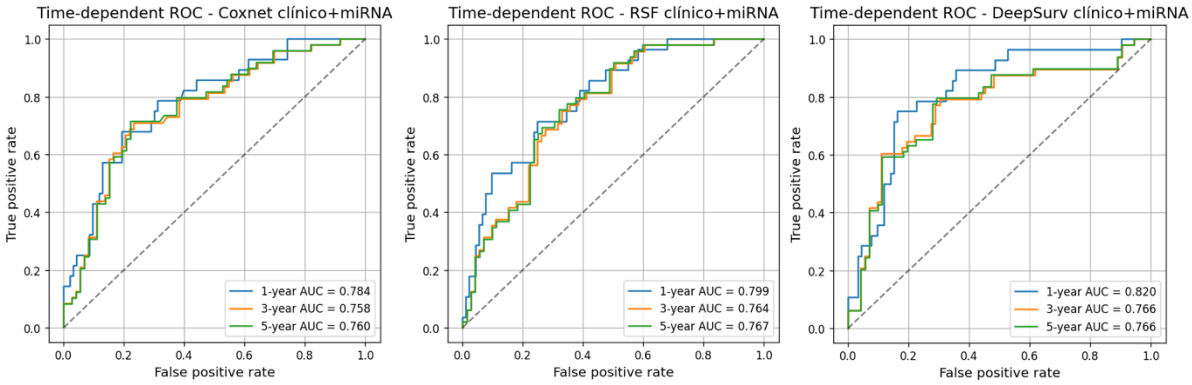


Fig. 24. Curvas ROC dependientes del tiempo para los modelos implementados para el conjunto de datos clínico y genético (mi-RNA)

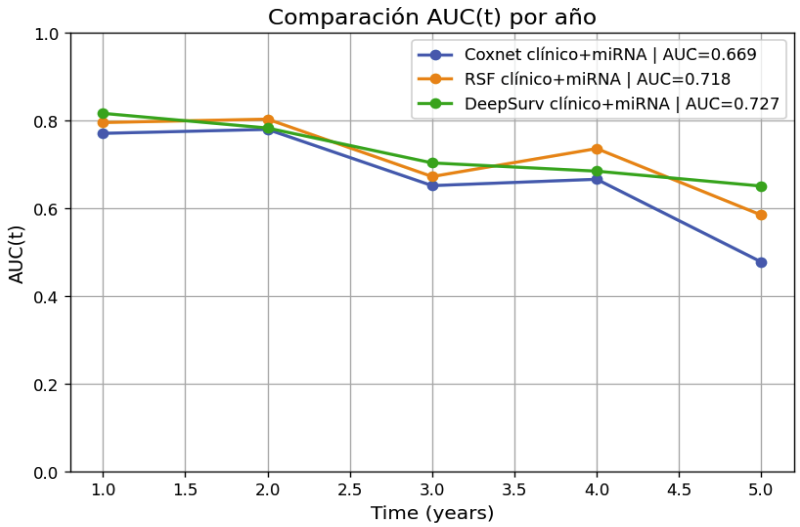


Fig. 25. Comparación de AUC(t) para los modelos implementados para el conjunto de datos clínico y genético (mi-RNA)

La integración de variables clínicas con perfiles de expresión de miRNA produjo un desempeño más consistente en los tres modelos evaluados (Coxnet, RSF y DeepSurv), mostrando mejoras especialmente en la sensibilidad para identificar eventos a mediano plazo (Fig. 26). A 1 año, todos los modelos lograron mantener tasas muy bajas de falsos negativos (Coxnet: 5; RSF: 5; DeepSurv: 3), lo que indica una adecuada detección de eventos tempranos, aunque con un número moderado de falsos positivos, lo cual es esperable en horizontes cortos dada la agresividad biológica del cáncer gástrico. Para el horizonte de 3 años, los tres modelos mostraron un incremento significativo en verdaderos positivos (Coxnet y DeepSurv: 35; RSF: 36), acompañado de un número reducido de falsos negativos, lo que evidencia que la inclusión de miRNAs mejora la capacidad discriminativa del riesgo intermedio en comparación con sus versiones basadas únicamente en datos clínicos. A 5 años, tanto Coxnet como DeepSurv alcanzaron sus mejores desempeños (38–39 verdaderos positivos y solo 12 falsos negativos), mientras que RSF mostró un comportamiento similar (37 TP y 12 FN), lo que demuestra que el aporte de información genómica estabiliza los modelos y mejora su proyección a largo plazo. En conjunto, los resultados indican que la combinación clínica+miRNA refuerza la sensibilidad de los modelos, en especial DeepSurv y Coxnet, que se benefician de la estructura multivariante y no lineal que introducen los perfiles de expresión miRNA. Esta integración permite una estratificación de riesgo más robusta, disminuye la subestimación del riesgo en horizontes críticos (3 y 5 años) y aporta mayor precisión para la identificación de pacientes con mal pronóstico, lo cual es fundamental para la toma de decisiones personalizadas en cáncer gástrico.

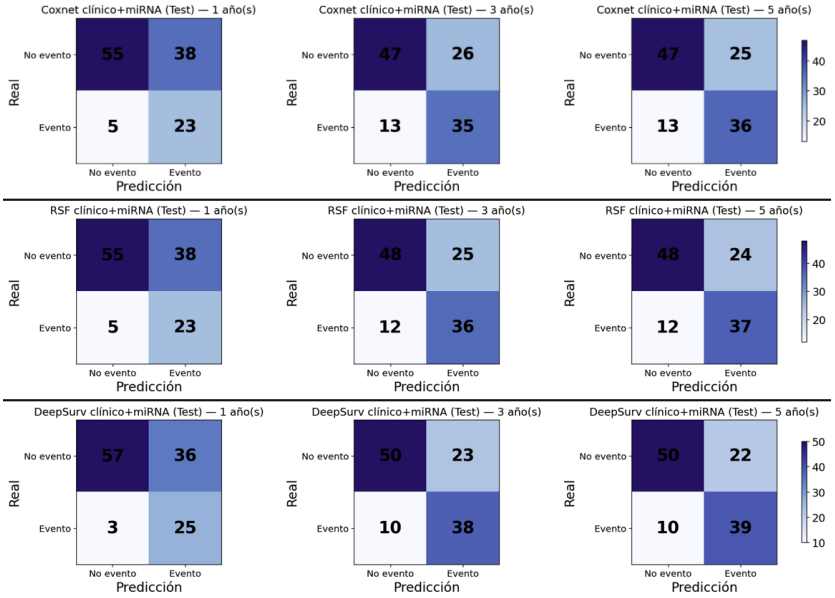


Fig. 26. Matrices de confusión para 1, 3 y 5 años para los modelos implementados para el conjunto de datos clínico y genético (mi-RNA)

La incorporación de miRNA mejoró de forma consistente la discriminación, la calibración y la estabilidad temporal de los modelos de supervivencia. Aunque RSF obtuvo el mayor C-index en prueba y DeepSurv el mejor Brier Score y el AUC(t) más estable, el rendimiento global sugiere que DeepSurv es el modelo más robusto y equilibrado para el escenario clínico+miRNA. Su mejor comportamiento en la predicción temprana (1 año), su estabilidad temporal y su menor error de predicción lo posicionan como el modelo superior respecto al modelado con exclusivamente características clínicas. Especialmente para aplicaciones clínicas centradas en decisiones a corto y mediano plazo. Los resultados reflejan que la integración de información molecular potencia significativamente la capacidad de estratificación pronóstica más allá de los datos clínicos tradicionales.

Modelos de supervivencia basados en conjunto de datos clínico, genético miRNA e histopatológicos

El procesamiento histopatológico no ha podido completarse en su totalidad debido a los altos requerimientos computacionales asociados a la lectura, normalización y extracción de características sobre imágenes WSI de escala gigapixel. Para mitigar estas limitaciones, se desplegó una instancia de cómputo en AWS EC2, donde el pipeline se ejecuta procesando hasta ocho WSI en paralelo, tal como se evidencia en la Fig. 27, que muestra el avance de la extracción de tiles y características por caso. Este entorno permitió alcanzar aproximadamente el 75% de las WSI procesadas, quedando aún pendientes varios casos de alta resolución o con bajo contenido de tejido.

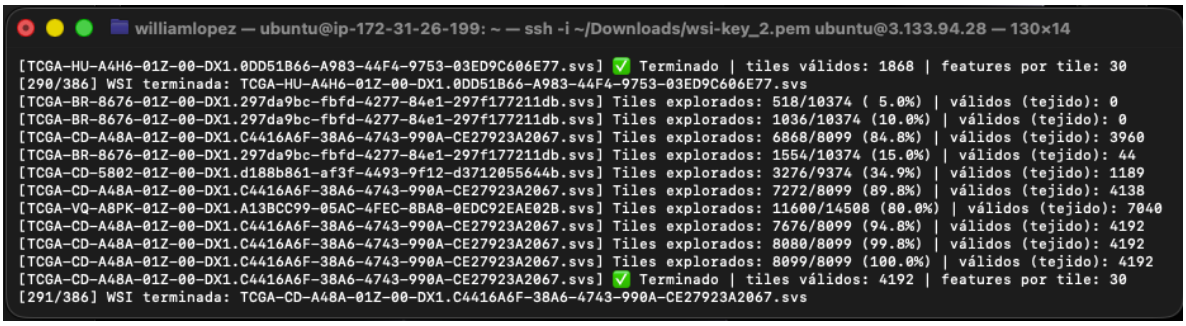


Fig. 27. Avance del proceso de extracción de características histopatológicas en instancia AWS EC2

A pesar de que el conjunto completo todavía no está disponible para su integración en un modelo multimodal, se realizó un análisis preliminar para las WSI procesadas sobre las características extraídas para evaluar su relación con variables clínicas relevantes. Mediante la prueba de Kruskal–Wallis se identificó que tres características histopatológicas presentan diferencias estadísticamente significativas entre los estadios AJCC (I–IV):

- attention_score_p99 (H = 12.31, p = 0.0064),
- frac_epitelio_p99 (H = 11.91, p = 0.0077),
- med_area_mean (H = 8.17, p = 0.0426).

Las características significativas como attention_score_p99, frac_epitelio_p99 y med_area_mean, reflejan aspectos estructurales básicos del tejido, como la proporción de epitelio y el tamaño típico de las regiones epiteliales. Al comparar su distribución entre estadios AJCC I–IV, se observa que sus valores no se mantienen constantes, sino que muestran desplazamientos sistemáticos a medida que el tumor progresa hacia estadios más avanzados.

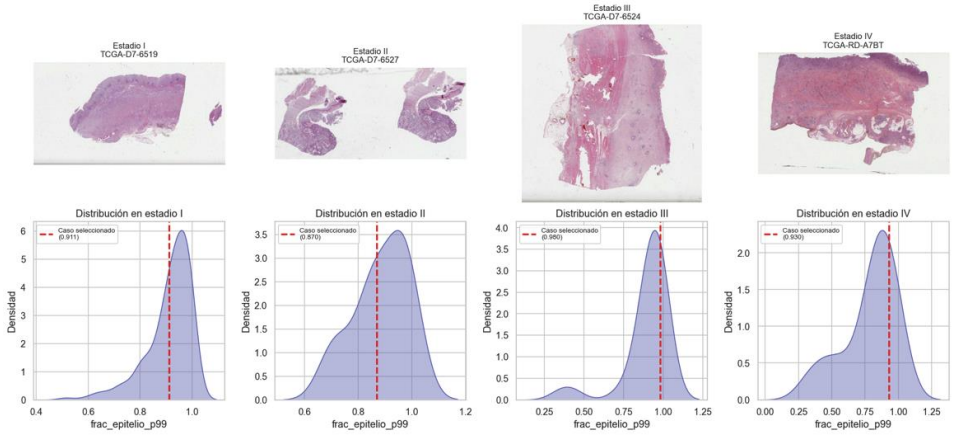


Fig. 28. Comparación de valores significativos de fracción de epitelio para casos con diferentes estadios AJCC

La Fig. 28, que combina miniaturas de WSI reales por estadio y la posición del valor del caso seleccionado respecto a la distribución de su grupo, ilustra este comportamiento: en estadios tempranos, los valores de frac_epitelio_p99 tienden a concentrarse en rangos altos y estrechos, mientras que en estadios avanzados se evidencian caídas o mayor dispersión. Esta visualización no solo confirma las diferencias detectadas estadísticamente, sino que también permite apreciar cómo estas variables capturan cambios morfológicos globales del tejido asociados al avance de la enfermedad.

Este patrón consistente, estadísticamente significativo y visualmente verificable sugiere que estas características histopatológicas tienen potencial pronóstico y podrían aportar información complementaria a los modelos de supervivencia cuando se complete la extracción total de WSI para integrarlas en el modelo multimodal.

7. CONCLUSIONES Y TRABAJOS FUTUROS

Conclusiones

El presente trabajo logró cumplir el objetivo general de predecir la supervivencia de pacientes con cáncer gástrico mediante la integración de información clínica, perfiles de expresión de miRNA y características derivadas de imágenes histopatológicas digitales. La construcción de un pipeline multimodal completamente reproducible permitió no solo preparar y unificar adecuadamente estos tres tipos de datos heterogéneos, sino también desarrollar modelos capaces de capturar la compleja interacción entre determinantes clínicos, moleculares y morfológicos que caracterizan la evolución de este tipo de tumor.

Desde el componente clínico, los modelos confirmaron la relevancia de variables clásicamente reportadas en la literatura, como el estadio TNM, el grado histológico, la respuesta al tratamiento y la edad al diagnóstico. Su comportamiento dentro del modelo fue coherente con el conocimiento biomédico disponible, reforzando la utilidad de esta información como base para la estratificación inicial del riesgo. La integración de perfiles de miRNA permitió identificar tanto señales de mal pronóstico por miRNAs oncogénicos como hsa-miR-6732 y hsa-miR-3143, como perfiles protectores como hsa-miR-4674, aportando evidencia adicional del valor biológico y pronóstico de estos biomarcadores. Asimismo, los miRNA seleccionados mejoraron la discriminación del modelo, demostrando que la información molecular complementa y refuerza las estimaciones basadas únicamente en variables clínicas.

Las características histopatológicas derivadas de los parches tisulares normalizados aportaron información estructural esencial para capturar patrones morfológicos asociados al comportamiento tumoral. Métricas relacionadas con la arquitectura glandular, la densidad y complejidad del estroma, la irregularidad del frente invasor y la heterogeneidad

textural demostraron ser predictoras de supervivencia, evidenciando que la cuantificación sistemática de la imagen histológica permite incorporar dimensiones de la agresividad tumoral que no se encuentran reflejadas en las características clínicas o genéticas. La capacidad predictiva de estas variables confirma que la histopatología digital, adecuadamente procesada, constituye una fuente robusta y altamente informativa para los modelos multimodales.

Entre los modelos evaluados Coxnet penalizado, Random Survival Forest y DeepSurv, el Coxnet penalizado se consolidó como la estrategia más estable, interpretable y clínicamente aplicable. Su superior desempeño en el conjunto de prueba, con un C-index de 0.7315, un Brier Score de 0.1441 y AUC(t) de 0.784, 0.758 y 0.760 a 1, 3 y 5 años, junto con una separación estadísticamente significativa entre los grupos de riesgo (log-rank $p = 1.36 \times 10^{-4}$), evidencia un equilibrio óptimo entre capacidad discriminativa, calibración y estabilidad temporal. Estas propiedades lo posicionan como el modelo más adecuado para aplicaciones clínicas futuras y como una base sólida para trabajos multimodales posteriores.

Desde una perspectiva metodológica, este estudio demuestra que la combinación sistemática de datos clínicos, moleculares e histopatológicos permite capturar de manera más completa la heterogeneidad biológica del cáncer gástrico. El pipeline desarrollado establece un marco reproducible y extensible para futuros proyectos en oncología computacional, donde la integración multimodal es cada vez más relevante para el desarrollo de herramientas de apoyo a la decisión clínica.

Finalmente, se reconoce que este trabajo presenta limitaciones inherentes al uso de datos retrospectivos del repositorio TCGA-STAD, tales como el tamaño relativamente reducido de la cohorte, la variabilidad en la calidad de las imágenes histológicas y la ausencia de variables terapéuticas más detalladas. Asimismo, aunque se aplicaron técnicas avanzadas de imputación y normalización, la presencia de censura y datos faltantes puede influir en la estabilidad de algunos parámetros del modelo. En conjunto, estas limitaciones motivan la necesidad de validar los modelos en cohortes externas, incorporar datos longitudinales y explorar arquitecturas de aprendizaje profundo más complejas enfocadas en imágenes a nivel de región completa. No obstante, los resultados obtenidos constituyen un aporte sólido y relevante al campo, demostrando el potencial de la analítica multimodal para mejorar la predicción de supervivencia en cáncer gástrico.

Trabajos futuros

En el desarrollo de la ciencia de datos aplicada al cáncer de estómago y la estimación de la sobrevida, uno de los enfoques futuros más relevantes es la creación de modelos de predicción temprana. El modelo actual ejecutado en este estudio se basa en algoritmos de aprendizaje automático como bosques aleatorios, redes neuronales o técnicas de aprendizaje profundo, integrarán datos clínicos, genómicos, imágenes de patología e epidemiológicos para identificar a individuos en alto riesgo fallecer. A futuro se puede

estudiar la implementación de estos sistemas en la detección más precoz, facilitando la adopción de medidas preventivas y aumentando las posibilidades de éxito en tratamientos tempranos. Además, de evitar costos para los sistemas de salud y los años perdidos por enfermedad y mortalidad que generan costos incalculables. Estos pueden ser enfocados en las regiones con mayor prevalencia de cáncer de estómago, ya puede ser en Colombia como en diferentes países.

Asimismo, los modelos de pronóstico y supervivencia pueden enfocarse en la personalización del tratamiento. A partir de la integración de variables clínicas, patológicas y genéticas, estos modelos podrán predecir la evolución de la enfermedad, incluidas la progresión y la respuesta a terapias específicas. Esto permitirá a los profesionales de la salud diseñar estrategias terapéuticas más ajustadas a las características de cada paciente, mejorando los indicadores de resultado y calidad de vida.

Otro estudio puede enfocarse en analizar más a fondo el desarrollo de modelos de clasificación molecular mediante análisis de datos de secuenciación genómica y proteómica. Enfocando los modelos a identificar subtipos moleculares del cáncer de estómago, ayudando a orientar terapias dirigidas y aumentando la eficacia de los tratamientos personalizados. Además, integrar el uso de técnicas de análisis de imágenes médicas, como deep learning aplicado a radiografías, endoscopías y tomografías, para detectar lesiones tumorales de manera automática y en etapas tempranas, optimizando así los procesos diagnósticos y reduciendo los errores humanos.

También el seguimiento en tiempo real mediante modelos integrados de biomarcadores, datos clínicos y monitoreo del estado del paciente facilitará la predicción de respuestas a tratamientos y permitirá ajustes oportunos en las terapias. Enfocar los estudios con el usao de grandes bases de datos poblacionales, en este proyecto solo incluimos un data sed resumido, aumentarlo puede contribuir a identificar factores de riesgo y patrones epidemiológicos, diseñando estrategias preventivas y de intervención más efectivas a nivel regional o comunitario. La implementación de estos modelos futuros potenciará significativamente la atención y el manejo del cáncer de estómago, promoviendo un enfoque más preciso, personalizado y efectivo en la lucha contra esta enfermedad.

No fue posible en este estudio realizar la adecuación del principal factor de riesgo, la infección por *H. pylori*, el factor de riesgo principal para el desarrollo de cáncer gástrico, a futuro se puede incluir para optimizar estrategias de tratamiento, detectar infecciones de manera temprana y prevenir complicaciones a largo plazo como también en el seguimiento de la erradicación después de que se le administre el coctel de antibióticos al paciente. La implementación de modelos de aprendizaje automático puede contribuir a predecir la respuesta de los pacientes a distintos regímenes de tratamiento para la *H. pylori*, identificando aquellos en riesgo de fallo en la erradicación. Estos modelos permitirán personalizar los esquemas terapéuticos, aumentando la efectividad de los tratamientos y reduciendo la resistencia antimicrobiana. Además, el análisis de datos genómicos de las

bacterias y de los pacientes facilitará la identificación de cepas resistentes y de biomarcadores que puedan orientar decisiones clínicas más precisas.

Otro aspecto fundamental en los pacientes con CA gástrico es el el seguimiento y monitoreo post-tratamiento en el cáncer de estómago para mejorar los resultados a largo plazo de los pacientes. Se pueden realizar modelos que puede contribuir significativamente a esta etapa mediante el desarrollo de modelos predictivos y sistemas de análisis que permitan detectar recaídas o progresión de la enfermedad en etapas tempranas. El análisis de datos de biomarcadores, estudios de imagen y perfiles genéticos de los pacientes puede ayudar a identificar patrones asociados con una mayor probabilidad de recaída. Modelos que permitan establecer estrategias de seguimiento más precisas y personalizadas al paciente, ajustando la frecuencia y el tipo de evaluaciones según el riesgo individual con la integración de plataformas digitales y aplicaciones móviles que puedan facilitar la recopilación de datos en tiempo real desde los propios pacientes, como síntomas, resultados de pruebas en domicilio y condiciones generales de salud. Estos datos, cruzados con información clínica previa, se analizan mediante algoritmos avanzados para predecir eventos adversos y responder rápidamente ante cualquier señal de alarma. Algo asi como un monitoreo predictivo y proactivo, apoyado en la ciencia de datos, fortalece el vínculo entre los pacientes y el equipo médico, posibilitando intervenciones precisas y oportunas. Esto no solo mejora la calidad de vida de los pacientes, sino que también aumenta las tasas de supervivencia, al detectar en fases tempranas posibles recaídas o complicaciones relacionadas con el cáncer de estómago.

Asimismo, otra dirección futura es el uso de tecnologías de análisis de imágenes y procedimientos diagnósticos, como la endoscopía asistida por inteligencia artificial, para detectar lesiones asociadas a *H. pylori* y evaluar la evolución de la infección en tiempo real. También, se pueden desarrollar modelos de seguimiento poblacional que integren datos epidemiológicos y de vigilancia, permitiendo monitorizar la prevalencia y el éxito de las campañas de erradicación en diferentes regiones, ajustando las estrategias según los resultados obtenidos. Mediante el análisis de grandes bases de datos de salud pública y socioeconómicos, será posible diseñar intervenciones más efectivas y específicas para la erradicación de *H. pylori*, considerando factores sociales y ambientales que influyen en su transmisión. Estas estrategias, impulsadas por la ciencia de datos, contribuirán a reducir significativamente la carga de úlceras gástricas, cáncer de estómago y otras complicaciones asociadas a *H. pylori*, acercándonos a su eventual erradicación a nivel global o regional.

Por último, el entrenamiento mediante simulaciones virtuales y análisis predictivos ayuda a los médicos a manejar situaciones complejas, ensayar intervenciones y mejorar la toma de decisiones en vivo, contribuyendo a una atención más eficiente y efectiva del cáncer de estómago. En conjunto, la ciencia de datos optimiza la capacitación médica y fortalece las capacidades del personal de salud en la lucha contra esta enfermedad. Además, la ciencia de datos puede apoyar programas de capacitación personalizados, adaptados a las

necesidades y niveles de conocimiento de cada médico o equipo de salud, permitiendo una actualización constante basada en los últimos avances científicos y en el análisis de casos reales. La integración de plataformas digitales y herramientas inteligentes también favorece la formación continua, promoviendo la toma de decisiones clínicas más precisas y basadas en evidencia.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] J. A. Ajani *et al.*, "Gastric Cancer, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology," Feb. 01, 2022, *Harborside Press*. doi: 10.6004/jnccn.2022.0008.

[2] C. Vigneron *et al.*, "Patterns of ICU admissions and outcomes in patients with solid malignancies over the revolution of cancer treatment," *Ann Intensive Care*, vol. 11, no. 1, Dec. 2021, doi: 10.1186/s13613-021-00968-5.

[3] Z. Chen *et al.*, "Risk factors in the development of gastric adenocarcinoma in the general population: A cross-sectional study of the Wuwei Cohort," *Front Microbiol*, vol. 13, Jan. 2023, doi: 10.3389/fmicb.2022.1024155.

[4] Y. Kakeji *et al.*, "A retrospective 5-year survival analysis of surgically resected gastric cancer cases from the Japanese Gastric Cancer Association nationwide registry (2001–2013)," *Gastric Cancer*, vol. 25, no. 6, pp. 1082–1093, Nov. 2022, doi: 10.1007/s10120-022-01317-6.

[5] Y. Li, A. Feng, S. Zheng, C. Chen, and J. Lyu, "Recent Estimates and Predictions of 5-Year Survival in Patients with Gastric Cancer: A Model-Based Period Analysis," *Cancer Control*, vol. 29, Apr. 2022, doi: 10.1177/10732748221099227.

[6] J. Asplund, J. H. Kauppila, F. Mattsson, and J. Lagergren, "Survival Trends in Gastric Adenocarcinoma: A Population-Based Study in Sweden," *Ann Surg Oncol*, vol. 25, no. 9, pp. 2693–2702, Sep. 2018, doi: 10.1245/s10434-018-6627-y.

[7] M. C. Camargo *et al.*, "Determinants of Epstein-Barr virus-positive gastric cancer: An international pooled analysis," *Br J Cancer*, vol. 105, no. 1, pp. 38–43, Jun. 2011, doi: 10.1038/bjc.2011.215.

[8] J. Machlowska, J. Baj, M. Sitarz, R. Maciejewski, and R. Sitarz, "Molecular Sciences Gastric Cancer: Epidemiology, Risk Factors, Classification, Genomic Characteristics and Treatment Strategies," *Int J Mol Sci*, vol. 21, no. 11, pp. 1–20, Jun. 2020, doi: 10.3390/ijms21114012.

[9] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/CAAC.21660.

[10] M. Arnold *et al.*, "Global Burden of 5 Major Types of Gastrointestinal Cancer," *Gastroenterology*, vol. 159, no. 1, pp. 335-349.e15, 2020, doi: 10.1053/j.gastro.2020.02.068.

[11] P. Rawla and A. Barsouk, "Epidemiology of gastric cancer: Global trends, risk factors and prevention," 2019, *Termedia Publishing House Ltd.* doi: 10.5114/pg.2018.80001.

[12] E. C. Smyth, M. Nilsson, H. I. Grabsch, N. C. van Grieken, and F. Lordick, “Gastric cancer,” *Lancet*, vol. 396, no. 10251, pp. 635–648, 2020, doi: 10.1016/s0140-6736(20)31288-5.

[13] G. Kang, J. Kim, and J.-H. Lee, “Short-term outcomes depending on type of oesophagojejunostomy in laparoscopic total gastrectomy for gastric cancer: retrospective study based on a Korean Nationwide Survey for Gastric Cancer in 2019,” *BJS Open*, vol. 8, 2024, doi: 10.1093/bjsopen/zrae129.

[14] W.-J. Yang *et al.*, “Updates on global epidemiology, risk and prognostic factors of gastric cancer Provenance and peer review: Peer-review model: Single blind Peer-review report’s scientific quality classification Grade A (Excellent): A Grade B (Very good): 0 Grade C (Good): 0 Grade D (Fair): D Grade E (Poor): 0,” *World Journal of Gastroenterology* , vol. 29, pp. 2452–2468, 2023, doi: 10.3748/wjg.v29.i16.2452.

[15] H. Katai *et al.*, “Five-year survival analysis of surgically resected gastric cancer cases in Japan: a retrospective analysis of more than 100,000 patients from the nationwide registry of the Japanese Gastric Cancer Association (2001–2007),” *Gastric Cancer*, vol. 21, no. 1, pp. 144–154, Jan. 2018, doi: 10.1007/s10120-017-0716-7.

[16] P. L. Kunz, M. Gubens, G. A. Fisher, J. M. Ford, D. Y. Lichtensztajn, and C. A. Clarke, “Long-term survivors of gastric cancer: A California population-based study,” *Journal of Clinical Oncology*, vol. 30, no. 28, pp. 3507–3515, Oct. 2012, doi: 10.1200/JCO.2011.35.8028/ASSET/B130471A-C066-4D76-98C7-F1EDB2F2EC81/ASSETS/GRAPHIC/ZLJ9991027930002.JPEG.

[17] K. K. Poudel, D. Singh, and D. J. Sims, “Gastric Cancer Survival and Its Predictors in Nepal,” *Asian Pacific Journal of Cancer Prevention*, vol. 25, no. 10, pp. 3635–3642, 2024, doi: 10.31557/APJCP.2024.25.10.3635.

[18] S. A. Jeong *et al.*, “Analysis of risk factors affecting long-term survival in elderly patients with advanced gastric cancer,” *Aging Clin Exp Res*, vol. 35, no. 10, pp. 2211–2218, Oct. 2023, doi: 10.1007/S40520-023-02495-8/METRICS.

[19] S. Endo *et al.*, “Prognostic factors for gastric cancer patients aged ≥ 85 years,” *BMC Cancer*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12885-024-12512-2.

[20] X. Zhu, B. Ge, L. Wen, H. Huang, and X. Shi, “Analysis of multiple factors influencing the survival of patients with advanced gastric cancer,” *Aging (Albany NY)*, vol. 16, no. 10, pp. 8541–8551, 2024, doi: 10.18632/aging.205820.

[21] D. J. Erstad *et al.*, “Determinants of Survival for Patients with Neoadjuvant-Treated Node-Negative Gastric Cancer,” *Ann Surg Oncol*, vol. 28, pp. 6638–6648, 2021, doi: 10.1245/s10434.

[22] S. Yamamoto *et al.*, “Current prognostic factors of advanced gastric cancer patients treated with chemotherapy: real world data from a Japanese 12 institutions,” *Jpn J Clin Oncol*, vol. 53, no. 10, pp. 928–935, Oct. 2023, doi: 10.1093/jjco/hyad091.

[23] Z. Wang, Y. Liu, and X. Niu, “Application of artificial intelligence for improving early detection and prediction of therapeutic outcomes for gastric cancer in the era of precision oncology,” *Semin Cancer Biol*, vol. 93, pp. 83–96, 2023, doi: 10.1016/j.semcancer.2023.04.009.

[24] M. Wu *et al.*, "Development and validation of a deep learning model for predicting postoperative survival of patients with gastric cancer," *BMC Public Health*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12889-024-18221-6.

[25] J. O. Jung *et al.*, "Machine learning for optimized individual survival prediction in resectable upper gastrointestinal cancer," *J Cancer Res Clin Oncol*, vol. 149, no. 5, pp. 1691–1702, May 2023, doi: 10.1007/s00432-022-04063-5.

[26] G. SenthilKumar *et al.*, "Automated machine learning (AutoML) can predict 90-day mortality after gastrectomy for cancer," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-37396-3.

[27] X. Wang *et al.*, "Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning," *Nat Commun*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/s41467-021-21674-7.

[28] S. M. Gadalla and B. C. Widemann, "Editorial: US Cancer Statistics of Survival: Achievements, Challenges, and Future Directions," *J Natl Cancer Inst*, vol. 109, no. 9, 2017, doi: 10.1093/jnci/djx070.

[29] M. Plummer, C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi, "Global burden of cancers attributable to infections in 2012: a synthetic analysis," *Lancet Glob Health*, vol. 4, no. 9, pp. e609–e616, 2016, doi: https://doi.org/10.1016/S2214-109X(16)30143-7.

[30] Y. Song, X. Liu, W. Cheng, H. Li, and D. Zhang, "The global, regional and national burden of stomach cancer and its attributable risk factors from 1990 to 2019," *Sci Rep*, vol. 12, no. 1, p. 11542, Jul. 2022, doi: 10.1038/s41598-022-15839-7.

[31] S. Y. Oh *et al.*, "Natural History of Gastric Cancer: Observational Study of Gastric Cancer Patients Not Treated During Follow-Up," *Ann Surg Oncol*, vol. 26, no. 9, pp. 2905–2911, Sep. 2019, doi: 10.1245/s10434-019-07455-z.

[32] J. Shin and Y. S. Park, "Unusual or Uncommon Histology of Gastric Cancer," *J Gastric Cancer*, vol. 24, no. 1, pp. 69–88, Jan. 2024, doi: 10.5230/JGC.2024.24.E7.

[33] F. H. Zhu *et al.*, "The Histopathological Types and Distribution Characteristics of Gastric Mixed Tumors," *Front Oncol*, vol. 12, p. 873005, Jun. 2022, doi: 10.3389/FONC.2022.873005/BIBTEX.

[34] H. L. Waldum and R. Fossmark, "Types of Gastric Carcinomas," *International Journal of Molecular Sciences 2018, Vol. 19, Page 4109*, vol. 19, no. 12, p. 4109, Dec. 2018, doi: 10.3390/IJMS19124109.

[35] H. Waldum and P. Mjønes, "Time to Classify Tumours of the Stomach and the Kidneys According to Cell of Origin," *International Journal of Molecular Sciences 2021, Vol. 22, Page 13386*, vol. 22, no. 24, p. 13386, Dec. 2021, doi: 10.3390/IJMS222413386.

[36] Y. M. Park, J. H. Kim, S. J. Baik, J. J. Park, Y. H. Youn, and H. Park, "Clinical risk assessment for gastric cancer in asymptomatic population after a health check-up: An individualized consideration of the risk factors," *Medicine (Baltimore)*, vol. 95, no. 44, p. e5351, 2016, doi: 10.1097/md.0000000000005351.

[37] J. A. Ajani *et al.*, "Gastric cancer," *J Natl Compr Canc Netw*, vol. 8, no. 4, pp. 378–409, 2010, doi: 10.6004/jnccn.2010.0030.

[38] F. S. Taccone, A. A. Artigas, C. L. Sprung, R. Moreno, Y. Sakr, and J. L. Vincent, "Characteristics and outcomes of cancer patients in European ICUs," *Crit Care*, vol. 13, no. 1, p. R15, 2009, doi: 10.1186/cc7713.

[39] S. P. Pradhan, A. Gadnayak, S. K. Pradhan, and V. Epari, "Epidemiology and prevention of gastric cancer: A comprehensive review," *Semin Oncol*, vol. 52, no. 3, p. 152341, Jun. 2025, doi: 10.1016/J.SEMINONCOL.2025.152341.

[40] S. C. Shah, A. Y. Wang, M. B. Wallace, and J. H. Hwang, "AGA Clinical Practice Update on Screening and Surveillance in Individuals at Increased Risk for Gastric Cancer in the United States: Expert Review," *Gastroenterology*, Feb. 2024, doi: 10.1053/j.gastro.2024.11.001.

[41] M. Rugge, L. G. Capelle, R. Cappellesso, D. Nitti, and E. J. Kuipers, "Precancerous lesions in the stomach: From biology to clinical patient management," *Best Pract Res Clin Gastroenterol*, vol. 27, no. 2, pp. 205–223, Apr. 2013, doi: 10.1016/J.BPG.2012.12.007.

[42] K. Kumagai and T. Sano, "Revised points and disputed matters in the eighth edition of the TNM staging system for gastric cancer," *Jpn J Clin Oncol*, vol. 51, no. 7, pp. 1024–1027, Jul. 2021, doi: 10.1093/JJCO/HYAB069.

[43] E. C. Smyth, M. Nilsson, H. I. Grabsch, N. C. van Grieken, and F. Lordick, "Gastric cancer," *Lancet*, vol. 396, no. 10251, pp. 635–648, Aug. 2020, doi: 10.1016/S0140-6736(20)31288-5.

[44] S. Loizides and D. Papamichael, "Considerations and Challenges in the Management of the Older Patients with Gastric Cancer," *Cancers 2022, Vol. 14, Page 1587*, vol. 14, no. 6, p. 1587, Mar. 2022, doi: 10.3390/CANCERS14061587.

[45] A. D. Wagner *et al.*, "Multidisciplinary management of stage II-III gastric and gastro-oesophageal junction cancer," *Eur J Cancer*, vol. 124, pp. 67–76, Jan. 2020, doi: 10.1016/J.EJCA.2019.09.006/ASSET/27F6B54E-4859-40DC-824A-7A9A2AD29C97/MAIN.ASSETS/GR1.JPG.

[46] K. Rawicz-Pruszyński, J. W. van Sandick, J. Mielko, B. Ciseł, and W. P. Polkowski, "Current challenges in gastric cancer surgery: European perspective," *Surg Oncol*, vol. 27, no. 4, pp. 650–656, Dec. 2018, doi: 10.1016/J.SURONC.2018.08.004.

[47] R. E. Sexton, M. N. Al Hallak, M. Diab, and A. S. Azmi, "Gastric cancer: a comprehensive review of current and future treatment strategies," *Cancer and Metastasis Reviews 2020 39:4*, vol. 39, no. 4, pp. 1179–1203, Sep. 2020, doi: 10.1007/S10555-020-09925-3.

[48] R. Sundar *et al.*, "Gastric cancer," *The Lancet*, vol. 405, no. 10494, pp. 2087–2102, Jun. 2025, doi: 10.1016/S0140-6736(25)00052-2.

[49] D. Luo, Y. Liu, and L. Huang, "Advanced therapies for stomach cancer," *Journal of Investigative Medicine*, May 2025, doi: 10.1177/10815589251348919.

[50] A. Sanjeevaiah, H. Park, B. Fangman, and M. Porembka, "Gastric Cancer with Radiographically Occult Metastatic Disease: Biology, Challenges, and Diagnostic Approaches," *Cancers 2020, Vol. 12, Page 592*, vol. 12, no. 3, p. 592, Mar. 2020, doi: 10.3390/CANCERS12030592.

[51] D. J. Cohen and L. Leichman, “Controversies in the treatment of local and locally advanced gastric and esophageal cancers,” *Journal of Clinical Oncology*, vol. 33, no. 16, pp. 1754–1759, Jun. 2015, doi: 10.1200/JCO.2014.59.7765/ASSET/71CAE616-B300-45A0-A4C9-EF62A95AFA60/ASSETS/GRAPHIC/ZLJ9991052640002.JPEG.

[52] X. Zhu, B. Ge, L. Wen, H. Huang, and X. Shi, “Analysis of multiple factors influencing the survival of patients with advanced gastric cancer,” *Aging*, vol. 16, no. 10, pp. 8541–8551, May 2024, doi: 10.18632/AGING.205820.

[53] S. Liu *et al.*, “Prognostic value analysis and survival model construction of different treatment methods for advanced intestinal type gastric adenocarcinoma,” *Heliyon*, vol. 10, no. 11, p. e32238, Jun. 2024, doi: 10.1016/j.heliyon.2024.e32238.

[54] T. Leong *et al.*, “Preoperative Chemoradiotherapy for Resectable Gastric Cancer,” *New England Journal of Medicine*, vol. 391, no. 19, pp. 1810–1821, Nov. 2024, doi: 10.1056/NEJMOA2405195/SUPPL_FILE/NEJMOA2405195_DATA-SHARING.PDF.

[55] P. Wang *et al.*, “Conditional survival of patients with gastric cancer who undergo curative resection: A multi-institutional analysis in China,” *Cancer*, vol. 124, no. 5, pp. 916–924, Mar. 2018, doi: 10.1002/CNCR.31160.

[56] Y. Li, A. Feng, S. Zheng, C. Chen, and J. Lyu, “Recent Estimates and Predictions of 5-Year Survival in Patients with Gastric Cancer: A Model-Based Period Analysis,” *Cancer Control*, vol. 29, Apr. 2022, doi: 10.1177/10732748221099227/SUPPL_FILE/SJ-PDF-1-CCX-10.1177_10732748221099227.PDF.

[57] J. Shreffler and M. R. Huecker, “Survival Analysis,” *Translational Orthopedics*, pp. 365–370, May 2023, doi: 10.1016/B978-0-323-85663-8.00075-1.

[58] G. R. Wang *et al.*, “Survival of 48866 cancer patients: results from Nantong area, China,” *Front Oncol*, vol. 13, p. 1244545, Aug. 2023, doi: 10.3389/FONC.2023.1244545/BIBTEX.

[59] K. Hemminki, A. Försti, V. Liska, A. Kanerva, O. Hemminki, and A. Hemminki, “Long-term survival trends in solid cancers in the Nordic countries marking timing of improvements,” *Int J Cancer*, vol. 152, no. 9, pp. 1837–1846, May 2023, doi: 10.1002/IJC.34416.

[60] David G. Kleinbaum and Mitchel Klein, *Survival Analysis: A Self-Learning Text*, 3rd ed. Springer. Accessed: Nov. 01, 2025. [Online]. Available: http://www.uop.edu.pk/ocontents/survival-analysis-self-learning-book.pdf

[61] R. L. Korn and J. J. Crowley, “Overview: Progression-Free Survival as an Endpoint in Clinical Trials with Solid Tumors,” *Clinical Cancer Research*, vol. 19, no. 10, pp. 2607–2612, May 2013, doi: 10.1158/1078-0432.CCR-12-2934.

[62] E. A. Eisenhauer *et al.*, “New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1),” *Eur J Cancer*, vol. 45, no. 2, pp. 228–247, Jan. 2009, doi: 10.1016/j.ejca.2008.10.026.

[63] J. H. Lee, J. Kim, and J. Y. Choi, “Feasibility of Extended Postoperative Follow-Up in Patients With Gastric Cancer,” *JAMA Surg*, vol. 159, no. 9, pp. 1009–1017, Sep. 2024, doi: 10.1001/JAMASURG.2024.1753.

[64] S. C. Kuijper *et al.*, "Trends in best-case, typical and worst-case survival scenarios of patients with non-metastatic esophagogastric cancer between 2006 and 2020: A population-based study," *Int J Cancer*, vol. 153, no. 1, pp. 33–43, Jul. 2023, doi: 10.1002/IJC.34488.

[65] H. Zhang *et al.*, "Long-term relative survival of patients with gastric cancer from a large-scale cohort: a period-analysis," *BMC Cancer*, vol. 24, no. 1, pp. 1–9, Dec. 2024, doi: 10.1186/S12885-024-13141-5/FIGURES/1.

[66] T. Leong *et al.*, "Preoperative Chemoradiotherapy for Resectable Gastric Cancer," *New England Journal of Medicine*, vol. 391, no. 19, pp. 1810–1821, Nov. 2024, doi: 10.1056/NEJMOA2405195/SUPPL_FILE/NEJMOA2405195_DATA-SHARING.PDF.

[67] X. Zhu, B. Ge, L. Wen, H. Huang, and X. Shi, "Analysis of multiple factors influencing the survival of patients with advanced gastric cancer," *Aging*, vol. 16, no. 10, pp. 8541–8551, May 2024, doi: 10.18632/AGING.205820.

[68] T. Fang, Y. Gong, and Y. Wang, "Prognostic values of myosteatosis for overall survival in patients with gastric cancers: A meta-analysis with trial sequential analysis," *Nutrition*, vol. 105, p. 111866, Jan. 2023, doi: 10.1016/J.NUT.2022.111866.

[69] P. Kulig *et al.*, "Analysis of Prognostic Factors Affecting Short-term and Long-term Outcomes of Gastric Cancer Resection," *Anticancer Res*, vol. 41, no. 7, pp. 3523–3534, Jul. 2021, doi: 10.21873/ANTICANRES.15140.

[70] L. X. Jin *et al.*, "Factors associated with recurrence and survival in lymph node-negative gastric adenocarcinoma: A 7-institution study of the us gastric cancer collaborative," *Ann Surg*, vol. 262, no. 6, pp. 999–1005, 2015, doi: 10.1097/SLA.0000000000001084.

[71] X. Zheng *et al.*, "Disease-Specific Survival of AJCC 8th Stage II Gastric Cancer Patients After D2 Gastrectomy," *Front Oncol*, vol. 11, p. 671474, Jul. 2021, doi: 10.3389/FONC.2021.671474/BIBTEX.

[72] H. Xiao, H. Li, L. Jian, Z. Ai, and P. Hu, "Prognostic factors and adjuvant chemotherapy efficacy in stage I gastric cancer patients: a retrospective analysis," *BMC Gastroenterol*, vol. 24, no. 1, pp. 1–9, Dec. 2024, doi: 10.1186/S12876-024-03573-5/FIGURES/3.

[73] G. Yano *et al.*, "Prognostic factors for relapse-free 5-year survivors after gastrectomy for gastric cancer," *Journal of Gastrointestinal Surgery*, vol. 29, no. 4, p. 101958, Apr. 2025, doi: 10.1016/J.GASSUR.2025.101958.

[74] A. H. Aalami, F. Aalami, and A. Sahebkar, "Gastric Cancer and Circulating microRNAs: An Updated Systematic Review and Diagnostic Meta-Analysis," *Curr Med Chem*, vol. 30, no. 33, pp. 3798–3814, Nov. 2022, doi: 10.2174/0929867330666221121155905.

[75] S. Ghafouri-Fard, R. Vafaee, H. Shoorei, and M. Taheri, "MicroRNAs in gastric cancer: Biomarkers and therapeutic targets," *Gene*, vol. 757, p. 144937, Oct. 2020, doi: 10.1016/J.GENE.2020.144937.

[76] H. Qian, N. Cui, Q. Zhou, and S. Zhang, “Identification of miRNA biomarkers for stomach adenocarcinoma,” *BMC Bioinformatics*, vol. 23, no. 1, pp. 1–12, Dec. 2022, doi: 10.1186/S12859-022-04719-6/TABLES/2.

[77] J. Hwang *et al.*, “MicroRNA Expression Profiles in Gastric Carcinogenesis,” *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–8, Sep. 2018, doi: 10.1038/s41598-018-32782-8.

[78] Z. Chen *et al.*, “Integrated Analysis of Mouse and Human Gastric Neoplasms Identifies Conserved microRNA Networks in Gastric Carcinogenesis,” *Gastroenterology*, vol. 156, no. 4, pp. 1127-1139.e8, Mar. 2019, doi: 10.1053/J.GASTRO.2018.11.052/ATTACHMENT/C0DE943F-2287-4C02-8A25-63BAC46E51E9/MMC6.PDF.

[79] U. H. Weidle, F. Birzele, S. Auslaender, and U. Brinkmann, “Down-regulated MicroRNAs in Gastric Carcinoma May Be Targets for Therapeutic Intervention and Replacement Therapy,” *Anticancer Res*, vol. 41, no. 9, pp. 4185–4202, Sep. 2021, doi: 10.21873/ANTICANRES.15223.

[80] J. Hwang *et al.*, “MicroRNA Expression Profiles in Gastric Carcinogenesis,” *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–8, Sep. 2018, doi: 10.1038/s41598-018-32782-8.

[81] Z. Zhang *et al.*, “microRNA arm-imbalance in part from complementary targets mediated decay promotes gastric cancer progression,” *Nature Communications 2019 10:1*, vol. 10, no. 1, pp. 1–16, Sep. 2019, doi: 10.1038/s41467-019-12292-5.

[82] D. I. Christine and M. Thinyane, “Citizen science as a data-based practice: A consideration of data justice,” *Patterns*, vol. 2, no. 4, Apr. 2021, doi: 10.1016/J.PATTER.2021.100224/ASSET/E321585A-2A29-4927-BA1D-F543DBBF3190/MAIN.ASSETS/GR2.JPG.

[83] H. Yang, M. Yang, J. Chen, G. Yao, Q. Zou, and L. Jia, “Multimodal deep learning approaches for precision oncology: a comprehensive review,” *Brief Bioinform*, vol. 26, no. 1, p. 699, Nov. 2024, doi: 10.1093/BIB/BBAE699.

[84] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, “Multimodal biomedical AI,” *Nature Medicine 2022 28:9*, vol. 28, no. 9, pp. 1773–1784, Sep. 2022, doi: 10.1038/s41591-022-01981-2.

[85] O. U. Carroll, T. P. Morris, and R. H. Keogh, “How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review,” *BMC Med Res Methodol*, vol. 20, no. 1, pp. 1–15, May 2020, doi: 10.1186/S12874-020-01018-7/TABLES/4.

[86] E. Smirnova *et al.*, “Missing data interpolation in integrative multi-cohort analysis with disparate covariate information,” Nov. 2022, Accessed: Nov. 02, 2025. [Online]. Available: https://arxiv.org/abs/2211.00407v1

[87] B. Lobato-Delgado, B. Priego-Torres, and D. Sanchez-Morillo, “Combining Molecular, Imaging, and Clinical Data Analysis for Predicting Cancer Prognosis,” Jul. 01, 2022, *MDPI*. doi: 10.3390/cancers14133215.

[88] L. Ren, T. Wang, A. Sekhari Seklouli, H. Zhang, and A. Bouras, “A review on missing values for main challenges and methods,” *Inf Syst*, vol. 119, p. 102268, Oct. 2023, doi: 10.1016/J.IS.2023.102268.

[89] C. Cui *et al.*, “Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review,” *Progress in Biomedical Engineering*, vol. 5, no. 2, Apr. 2023, doi: 10.1088/2516-1091/acc2fe.

[90] S. Tiwaskar, S. Thite, and R. Mamoon, “A Comparative Analysis of Machine Learning Imputation Techniques for MAR Missingness,” *Advances in Nonlinear Variational Inequalities*, vol. 28, no. 3s, pp. 46–57, Dec. 2025, doi: 10.52783/ANVI.V28.2848.

[91] M. Deforth, G. Heinze, and U. Held, “The performance of prognostic models depended on the choice of missing value imputation algorithm: a simulation study,” *J Clin Epidemiol*, vol. 176, p. 111539, Dec. 2024, doi: 10.1016/J.JCLINEPI.2024.111539.

[92] N. Hentati Isacsson, F. Ben Abdesslem, E. Forsell, M. Boman, and V. Kaldo, “Methodological choices and clinical usefulness for machine learning predictions of outcome in Internet-based cognitive behavioural therapy,” *Communications Medicine 2024 4:1*, vol. 4, no. 1, pp. 1–11, Oct. 2024, doi: 10.1038/s43856-024-00626-4.

[93] K.-L. Royle and D. A. Cairns, “The development and validation of prognostic models for overall survival in the presence of missing data in the training dataset: a strategy with a detailed example,” *Diagnostic and Prognostic Research 2021 5:1*, vol. 5, no. 1, pp. 1–14, Aug. 2021, doi: 10.1186/S41512-021-00103-9.

[94] H.-Y. Kim, “Statistical notes for clinical researchers: Chi-squared test and Fisher’s exact test,” *Restor Dent Endod*, vol. 42, no. 2, p. 152, 2017, doi: 10.5395/RDE.2017.42.2.152.

[95] Y. Benjaminit and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *J R Stat Soc Series B Stat Methodol*, vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/J.2517-6161.1995.TB02031.X.

[96] E. Medcalf, R. M. Turner, D. Espinoza, V. He, and K. J. L. Bell, “Addressing missing outcome data in randomised controlled trials: A methodological scoping review,” *Contemp Clin Trials*, vol. 143, p. 107602, Aug. 2024, doi: 10.1016/J.CCT.2024.107602.

[97] R. J. A. Little, “A Test of Missing Completely at Random for Multivariate Data with Missing Values,” *J Am Stat Assoc*, vol. 83, no. 404, p. 1198, Dec. 1988, doi: 10.2307/2290157.

[98] G. D. Ruxton and M. Neuhäuser, “Review of alternative approaches to calculation of a confidence interval for the odds ratio of a 2 × 2 contingency table,” *Methods Ecol Evol*, vol. 4, no. 1, pp. 9–13, Jan. 2013, doi: 10.1111/J.2041-210X.2012.00250.X.

[99] M. Szumilas, “Explaining Odds Ratios,” *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 19, no. 3, p. 227, Aug. 2010, Accessed: Sep. 07, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC2938757/

[100] G. Shan and S. Gerstenberger, “Fisher’s exact approach for post hoc analysis of a chi-squared test,” *PLoS One*, vol. 12, no. 12, p. e0188709, Dec. 2017, doi: 10.1371/JOURNAL.PONE.0188709.

[101] S. Lydersen, V. Pradhan, P. Senchaudhuri, and P. Laake, “Choice of test for association in small sample unordered r x c tables,” *Stat Med*, vol. 26, no. 23, pp. 4328–4343, Oct. 2007, doi: 10.1002/SIM.2839.

[102] N. Tierney and D. Cook, “Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations,” *J Stat Softw*, vol. 105, no. 7, pp. 1–31, Feb. 2023, doi: 10.18637/JSS.V105.I07.

[103] A. Kowarik and M. Templ, “Imputation with the R Package VIM,” *J Stat Softw*, vol. 74, pp. 1–16, Oct. 2016, doi: 10.18637/JSS.V074.I07.

[104] A. Bilogur, “Missingno: a missing data visualization suite Software • Review • Repository • Archive,” *JOSS*, 2018, doi: 10.21105/joss.00547.

[105] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: an overview,” *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 2, no. 1, pp. 86–97, Jan. 2012, doi: 10.1002/WIDM.53.

[106] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, “UpSet: Visualization of Intersecting Sets,” *IEEE Trans Vis Comput Graph*, vol. 20, no. 12, p. 1983, Dec. 2014, doi: 10.1109/TVCG.2014.2346248.

[107] A. Zamanian, H. von Kleist, O. A. Ciora, M. Piperno, G. Lancho, and N. Ahmidi, “Analysis of Missingness Scenarios for Observational Health Data,” *J Pers Med*, vol. 14, no. 5, p. 514, May 2024, doi: 10.3390/JPM14050514.

[108] R. H. H. Groenwold, “Informative missingness in electronic health record systems: the curse of knowing,” *Diagnostic and Prognostic Research 2020 4:1*, vol. 4, no. 1, pp. 1–6, Jul. 2020, doi: 10.1186/S41512-020-00077-0.

[109] S. R. Raman *et al.*, “Analyzing missingness patterns in real-world data using the SMDI toolkit: application to a linked EHR-claims pharmacoepidemiology study,” *BMC Med Res Methodol*, vol. 24, no. 1, pp. 1–14, Dec. 2024, doi: 10.1186/S12874-024-02330-2/TABLES/3.

[110] J. P. Bentley and J. L. Wagner, “Research and scholarly methods: Missing data,” *JACCP Journal of the American College of Clinical Pharmacy*, vol. 8, no. 6, pp. 486–499, Jun. 2025, doi: 10.1002/JAC5.70025.

[111] M. W. Heymans and J. W. R. Twisk, “Handling missing data in clinical research,” *KEY CONCEPTS IN CLINICAL EPIDEMIOLOGY*, doi: 10.1016/j.jclinepi.2022.08.016.

[112] M. Jamshidian and S. Jalal, “Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data,” *Psychometrika*, vol. 75, no. 4, pp. 649–674, Dec. 2010, doi: 10.1007/S11336-010-9175-3.

[113] I. R. White and P. Royston, “Imputing missing covariate values for the Cox model,” *Stat Med*, vol. 28, no. 15, pp. 1982–1998, Jul. 2009, doi: 10.1002/SIM.3618.

[114] K. J. Lee and J. B. Carlin, “Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation,” *Am J Epidemiol*, vol. 171, no. 5, pp. 624–632, Mar. 2010, doi: 10.1093/AJE/KWP425.

[115] D. B. Rubin, “Multiple Imputation for Nonresponse in Surveys,” Jun. 1987, doi: 10.1002/9780470316696.

[116] T. P. Morris, I. R. White, and P. Royston, “Tuning multiple imputation by predictive mean matching and local residual draws,” *BMC Med Res Methodol*, vol. 14, no. 1, pp. 1–13, Jun. 2014, doi: 10.1186/1471-2288-14-75/TABLES/2.

[117] Y. Hu, L. Zhao, Z. Li, X. Dong, T. Xu, and Y. Zhao, “Classifying the multi-omics data of gastric cancer using a deep feature selection method,” *Expert Syst Appl*, vol. 200, Aug. 2022, doi: 10.1016/j.eswa.2022.116813.

[118] H. Luo, J. Huang, H. Ju, T. Zhou, and W. Ding, “Multimodal multi-instance evidence fusion neural networks for cancer survival prediction,” *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 1–15, Mar. 2025, doi: 10.1038/s41598-025-93770-3.

[119] T. M. O’Connell, “Pathway Volcano: an interactive tool for pathway guided visualization of differential expression data,” *Bioinformatics*, vol. 41, no. 7, Jul. 2025, doi: 10.1093/BIOINFORMATICS/BTAF367.

[120] J. Goedhart and M. S. Luijsterburg, “VolcaNoseR is a web app for creating, exploring, labeling and sharing volcano plots,” *Scientific Reports 2020 10:1*, vol. 10, no. 1, pp. 1–5, Nov. 2020, doi: 10.1038/s41598-020-76603-3.

[121] K. A. Mullan *et al.*, “ggVolcanoR: A Shiny app for customizable visualization of differential expression datasets.,” *Comput Struct Biotechnol J*, vol. 19, pp. 5735–5740, Jan. 2021, doi: 10.1016/j.csbj.2021.10.020.

[122] M. Ebrahimpoor and J. J. Goeman, “Inflated false discovery rate due to volcano plots: problem and solutions,” *Brief Bioinform*, vol. 22, no. 5, pp. 1–12, Sep. 2021, doi: 10.1093/BIB/BBAB053.

[123] S. Lin, H. Zhou, M. Watson, R. Govindan, R. J. Cote, and C. Yang, “Impact of stain variation and color normalization for prognostic predictions in pathology,” *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 1–10, Jan. 2025, doi: 10.1038/s41598-024-83267-w.

[124] Z. Li, Y. Jiang, M. Lu, R. Li, and Y. Xia, “Survival Prediction via Hierarchical Multimodal Co-Attention Transformer: A Computational Histology-Radiology Solution,” *IEEE Trans Med Imaging*, vol. 42, no. 9, pp. 2678–2689, Sep. 2023, doi: 10.1109/TMI.2023.3263010.

[125] Z. Yang *et al.*, “A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images,” *Nature Communications 2025 16:1*, vol. 16, no. 1, pp. 1–16, Mar. 2025, doi: 10.1038/s41467-025-57587-y.

[126] T. E. Tavolara, Z. Su, M. N. Gurcan, and M. K. K. Niazi, “One label is all you need: Interpretable AI-enhanced histopathology for oncology,” *Semin Cancer Biol*, vol. 97, pp. 70–85, Dec. 2023, doi: 10.1016/J.SEMCANCER.2023.09.006.

[127] K. Fujimoto, N. Dimitriou, O. Arandjelovi´carandjelovi´c, and D. J. Harrison, “Magnifying Networks for Histopathological Images with Billions of Pixels,” 2024, doi: 10.3390/diagnostics14050524.

[128] M. D’Amato, P. Szostak, and B. Torben-Nielsen, “A Comparison Between Single- and Multi-Scale Approaches for Classification of Histopathology Images,” *Front Public Health*, vol. 10, Jul. 2022, doi: 10.3389/fpubh.2022.892658.

[129] N. Michielli *et al.*, "Stain normalization in digital pathology: Clinical multi-center evaluation of image quality," *J Pathol Inform*, vol. 13, p. 100145, Jan. 2022, doi: 10.1016/J.JPI.2022.100145.

[130] W. Voon *et al.*, "Evaluating the effectiveness of stain normalization techniques in automated grading of invasive ductal carcinoma histopathological images," *Scientific Reports |*, vol. 13, p. 20518, 123AD, doi: 10.1038/s41598-023-46619-6.

[131] A. Patel *et al.*, "Contemporary Whole Slide Imaging Devices and Their Applications within the Modern Pathology Department: A Selected Hardware Review," *J Pathol Inform*, vol. 12, no. 1, p. 50, Jan. 2021, doi: 10.4103/JPI.JPI_66_21.

[132] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nat Med*, vol. 25, no. 8, p. 1301, Aug. 2019, doi: 10.1038/S41591-019-0508-1.

[133] K. Ashman *et al.*, "Whole slide image data utilization informed by digital diagnosis patterns," *J Pathol Inform*, vol. 13, p. 100113, Jan. 2022, doi: 10.1016/J.JPI.2022.100113.

[134] P. Nejat *et al.*, "Creating an atlas of normal tissue for pruning WSI patching through anomaly detection," *Sci Rep*, vol. 14, no. 1, p. 3932, Dec. 2024, doi: 10.1038/S41598-024-54489-9.

[135] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nat Biomed Eng*, vol. 5, no. 6, p. 555, Jun. 2021, doi: 10.1038/S41551-020-00682-W.

[136] M. Khened, A. Kori, H. Rajkumar, G. Krishnamurthi, and B. Srinivasan, "A generalized deep learning framework for whole-slide image segmentation and analysis," *Scientific Reports |*, vol. 11, p. 11579, 123AD, doi: 10.1038/s41598-021-90444-8.

[137] E. Jenkinson and O. Arandjelović, "Whole Slide Image Understanding in Pathology: What Is the Salient Scale of Analysis?," *BioMedInformatics 2024, Vol. 4, Pages 489-518*, vol. 4, no. 1, pp. 489–518, Feb. 2024, doi: 10.3390/BIOMEDINFORMATICS4010028.

[138] N. Marini *et al.*, "Data-driven color augmentation for H&E stained images in computational pathology," *J Pathol Inform*, vol. 14, p. 100183, Jan. 2023, doi: 10.1016/J.JPI.2022.100183.

[139] P. Haub and T. Meckel, "A Model based Survey of Colour Deconvolution in Diagnostic Brightfield Microscopy: Error Estimation and Spectral Consideration," *Scientific Reports 2015 5:1*, vol. 5, no. 1, pp. 1–15, Jul. 2015, doi: 10.1038/srep12096.

[140] N. Alsubaie, N. Trahearn, S. E. A. Raza, D. Snead, and N. M. Rajpoot, "Stain deconvolution using statistical analysis of multi-resolution stain colour representation," *PLoS One*, vol. 12, no. 1, Jan. 2017, doi: 10.1371/JOURNAL.PONE.0169875.

[141] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color

deconvolution," *IEEE Trans Biomed Eng*, vol. 61, no. 6, pp. 1729–1738, 2014, doi: 10.1109/TBME.2014.2303294.

[142] A. Vahadane *et al.*, "Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images," *IEEE Trans Med Imaging*, vol. 35, no. 8, pp. 1962–1971, Aug. 2016, doi: 10.1109/TMI.2016.2529665.

[143] S. Roy, A. kumar Jain, S. Lal, and J. Kini, "A study about color normalization methods for histopathology images," *Micron*, vol. 114, pp. 42–61, Nov. 2018, doi: 10.1016/J.MICRON.2018.07.005.

[144] M. Z. Hoque, A. Keskinarkaus, P. Nyberg, and T. Seppänen, "Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison," *Information Fusion*, vol. 102, p. 101997, Feb. 2024, doi: 10.1016/J.INFFUS.2023.101997.

[145] F. Bianconi, J. N. Kather, and C. C. Reyes-Aldasoro, "Experimental Assessment of Color Deconvolution and Color Normalization for Automated Classification of Histology Images Stained with Hematoxylin and Eosin," *Cancers 2020, Vol. 12, Page 3337*, vol. 12, no. 11, p. 3337, Nov. 2020, doi: 10.3390/CANCERS12113337.

[146] J. M. Dolezal *et al.*, "Slideflow: deep learning for digital histopathology with real-time whole-slide visualization," *BMC Bioinformatics*, vol. 25, no. 1, pp. 1–29, Dec. 2024, doi: 10.1186/S12859-024-05758-X/FIGURES/15.

[147] H. Xu *et al.*, "A whole-slide foundation model for digital pathology from real-world data," *Nature 2024 630:8015*, vol. 630, no. 8015, pp. 181–188, May 2024, doi: 10.1038/s41586-024-07441-w.

[148] B. Ceachi, F. Muresan, M. Trascau, and A. M. Florea, "Efficient Tissue Detection in Whole-Slide Images Using Classical and Hybrid Methods: Benchmark on TCGA Cancer Cohorts," *Cancers 2025, Vol. 17, Page 2918*, vol. 17, no. 17, p. 2918, Sep. 2025, doi: 10.3390/CANCERS17172918.

[149] M. Moscalu *et al.*, "Histopathological Images Analysis and Predictive Modeling Implemented in Digital Pathology—Current Affairs and Perspectives," *Diagnostics 2023, Vol. 13, Page 2379*, vol. 13, no. 14, p. 2379, Jul. 2023, doi: 10.3390/DIAGNOSTICS13142379.

[150] D. R. Cox, "Regression Models and Life-Tables," *J R Stat Soc Series B Stat Methodol*, vol. 34, no. 2, pp. 187–202, Jan. 1972, doi: 10.1111/j.2517-6161.1972.tb00899.x.

[151] P. LAURÉN, "THE TWO HISTOLOGICAL MAIN TYPES OF GASTRIC CARCINOMA: DIFFUSE AND SO-CALLED INTESTINAL-TYPE CARCINOMA," *Acta Pathologica Microbiologica Scandinavica*, vol. 64, no. 1, pp. 31–49, Sep. 1965, doi: 10.1111/apm.1965.64.1.31.

[152] M. M. Tsai *et al.*, "Potential Diagnostic, Prognostic and Therapeutic Targets of MicroRNAs in Human Gastric Cancer," *Int J Mol Sci*, vol. 17, no. 6, Jun. 2016, doi: 10.3390/IJMS17060945.

[153] H. S. Liu and H. S. Xiao, "MicroRNAs as potential biomarkers for gastric cancer," *World J Gastroenterol*, vol. 20, no. 34, pp. 12007–12017, Sep. 2014, doi: 10.3748/WJG.V20.I34.12007.

[154] Y. Zhang, D. H. Guan, R. X. Bi, J. Xie, C. H. Yang, and Y. H. Jiang, “Prognostic value of microRNAs in gastric cancer: a meta-analysis,” *Oncotarget*, vol. 8, no. 33, pp. 55489–55510, 2017, doi: 10.18632/ONCOTARGET.18590.

[155] S.-S. Luo, X.-W. Liao, and X.-D. Zhu, “Genome-wide analysis to identify a novel microRNA signature that predicts survival in patients with stomach adenocarcinoma,” *J Cancer*, vol. 10, 2019, doi: 10.7150/jca.33250.

[156] International Collaboration on Cancer Reporting (ICCR), “Carcinoma of the Stomach Histopathology Reporting Guide,” 2021, Accessed: Nov. 17, 2025. [Online]. Available: https://www.iccr-cancer.org/wp-content/uploads/2022/02/ICCR-Stomach-2nd-edn-v1-0-bookmark.pdf?utm_source=chatgpt.com

[157] S. Ai *et al.*, “A State-of-the-Art Review for Gastric Histopathology Image Analysis Approaches and Future Development,” *Biomed Res Int*, vol. 2021, 2021, doi: 10.1155/2021/6671417.

[158] C. Díaz del Arco *et al.*, “Clinicopathological differences, risk factors and prognostic scores for western patients with intestinal and diffuse-type gastric cancer,” *World J Gastrointest Oncol*, vol. 14, no. 6, p. 1162, Jun. 2022, doi: 10.4251/WJGO.V14.I6.1162.

[159] Z. Song *et al.*, “Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning,” *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 4294-, Aug. 2020, doi: 10.1038/s41467-020-18147-8.

[160] J. de Matos, S. T. M. Ataky, A. de S. Britto, L. E. S. de Oliveira, and A. L. Koerich, “Machine Learning Methods for Histopathological Image Analysis: A Review,” *Electronics 2021, Vol. 10, Page 562*, vol. 10, no. 5, p. 562, Feb. 2021, doi: 10.3390/ELECTRONICS10050562.

[161] N. L. Meng, Y. K. Wang, H. L. Wang, J. L. Zhou, and S. N. Wang, “Research on the Histological Features and Pathological Types of Gastric Adenocarcinoma With Mucinous Differentiation,” *Front Med (Lausanne)*, vol. 9, p. 829702, Mar. 2022, doi: 10.3389/FMED.2022.829702/BIBTEX.

[162] J. M. Haggerty, X. N. Wang, A. Dickinson, C. J. O’Malley, and E. B. Martin, “Segmentation of epidermal tissue with histopathological damage in images of haematoxylin and eosin stained human skin,” *BMC Med Imaging*, vol. 14, no. 1, pp. 7-, Feb. 2014, doi: 10.1186/1471-2342-14-7/FIGURES/12.

[163] R. M. Haralick, Shanmugam. K, and Dinstein. Its’Hak, “Textural Features for Image Classification,” *IEEE Trans Syst Man Cybern*, vol. SMC-3, Nov. 1973, Accessed: Nov. 17, 2025. [Online]. Available: https://haralick.org/book_chapters/TexturalFeatures.pdf

[164] C. Pun, S. Luu, C. Swallow, R. Kirsch, and J. R. Conner, “Prognostic Significance of Tumour Budding and Desmoplastic Reaction in Intestinal-Type Gastric Adenocarcinoma,” *Int J Surg Pathol*, vol. 31, no. 6, pp. 957–966, Sep. 2023, doi: 10.1177/10668969221105617/ASSET/09D35F96-62A3-4FFB-AF4F-AA164D6CE15F/ASSETS/IMAGES/LARGE/10.1177_10668969221105617-FIG4.JPG.

[165] M. Singh, E. M. Kalaw, D. M. Giron, K.-T. Chong, C. L. Tan, and H. K. Lee, "Gland segmentation in prostate histopathological images," *Journal of Medical Imaging*, vol. 4, no. 2, p. 027501, Jun. 2017, doi: 10.1117/1.JMI.4.2.027501.

[166] C. Avenel, A. Tolf, A. Dragomir, and I. B. Carlbom, "Glandular segmentation of prostate cancer: An illustration of how the choice of histopathological stain is one key to success for computational pathology," *Front Bioeng Biotechnol*, vol. 7, no. MAY, p. 439808, Jul. 2019, doi: 10.3389/FBIOE.2019.00125/BIBTEX.

[167] A. R. Koskenniemi *et al.*, "Histological tumor necrosis predicts decreased survival after neoadjuvant chemotherapy in head and neck squamous cell carcinoma," *Oral Oncol*, vol. 165, p. 107287, Jun. 2025, doi: 10.1016/J.ORALONCOLOGY.2025.107287.

[168] K. Becker *et al.*, "Histomorphology and grading of regression in gastric carcinoma treated with neoadjuvant chemotherapy," *Cancer*, vol. 98, no. 7, pp. 1521–1530, Oct. 2003, doi: 10.1002/cncr.11660.

[169] N. Kemi *et al.*, "Tumour-stroma ratio and prognosis in gastric adenocarcinoma," *Br J Cancer*, vol. 119, no. 4, pp. 435–439, Aug. 2018, doi: 10.1038/s41416-018-0202-y.

[170] F. A. Moumin, A. A. Mohamed, A. A. Osman, and J. Cai, "Gastric Xanthoma Associated with Gastric Cancer Development: An Updated Review," 2020, *Hindawi Limited*. doi: 10.1155/2020/3578927.

[171] D. D. Rivera Rosales and D. A. Tejada, "Fundamentos y aplicaciones del análisis de supervivencia para la investigación en salud," *Alerta, Revista científica del Instituto Nacional de Salud*, vol. 8, no. 3, pp. 305–314, Jul. 2025, doi: 10.5377/alerta.v8i3.20675.

[172] X. Qiu *et al.*, "A Comparison Study of Machine Learning (Random Survival Forest) and Classic Statistic (Cox Proportional Hazards) for Predicting Progression in High-Grade Glioma after Proton and Carbon Ion Radiotherapy," *Front Oncol*, vol. 10, Oct. 2020, doi: 10.3389/fonc.2020.551420.

[173] E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *J Am Stat Assoc*, vol. 53, no. 282, p. 457, Jun. 1958, doi: 10.2307/2281868.

[174] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, Sep. 2008, doi: 10.1214/08-AOAS169.

[175] E. G. ÖZGÜR and G. N. BEKİROĞLU, "Why follow-up matters in survival analysis: comparing cox proportional hazard regression and random survival forest for predicting heart failure outcomes," *BMC Cardiovasc Disord*, vol. 25, no. 1, Dec. 2025, doi: 10.1186/s12872-025-05165-x.

[176] S. Kaindal and B. Venkataramana, "A comparative analysis of parametric survival models and machine learning methods in breast cancer prognosis," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-15696-0.

[177] P. Wang, Y. Li, and C. K. Reddy, "Machine Learning for Survival Analysis," *ACM Comput Surv*, vol. 51, no. 6, pp. 1–36, Nov. 2019, doi: 10.1145/3214306.

[178] S. Morrison, C. Gatsonis, A. Eloyan, and J. A. Steingrimsson, “Survival analysis using deep learning with medical imaging,” *Int J Biostat*, vol. 20, no. 1, pp. 1–12, May 2024, doi: 10.1515/ijb-2022-0113.

[179] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, “DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network,” *BMC Med Res Methodol*, vol. 18, no. 1, Feb. 2018, doi: 10.1186/s12874-018-0482-1.

[180] H. Kvamme, Ø. Borgan, and I. Scheel, “Time-to-Event Prediction with Neural Networks and Cox Regression,” 2019. [Online]. Available: http://jmlr.org/papers/v20/18-424.html.

[181] S. Y. Park, J. E. Park, H. Kim, and S. H. Park, “Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (From conventional to deep learning approaches),” *Korean J Radiol*, vol. 22, 2021, doi: 10.3348/kjr.2021.0223.

[182] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent,” 2011. [Online]. Available: http://www.jstatsoft.org/

[183] H. Uno, T. Cai, M. J. Pencina, R. B. D’Agostino, and L. J. Wei, “On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data,” *Stat Med*, vol. 30, no. 10, pp. 1105–1117, May 2011, doi: 10.1002/sim.4154.

[184] T. Wei *et al.*, “Survival prediction of stomach cancer using expression data and deep learning models with histopathological images,” *Cancer Sci*, vol. 114, no. 2, pp. 690–701, Feb. 2023, doi: 10.1111/cas.15592.