Facultad de Ciencias, Grado en Estadística

Trabajo de Fin de Grado

Análisis y caracterización de pacientes con Leucemia Mieloide Aguda

Analysis and characterization of patients with Acute Myeloid Leukemia

Autora:

Aroa Palomo Vadillo

Tutores:

Dr. José Manuel Sánchez Santos

Ángela Villaverde Ramiro

Julio, 2023

Facultad de Ciencias, Grado en Estadística

Trabajo de Fin de Grado

Análisis y caracterización de pacientes con leucemia mieloide aguda

Analysis and characterization of patients with Acute Myeloid Leukemia

Autora:

Aroa Palomo Vadillo

Tutores:

Dr. José Manuel Sánchez Santos

Ángela Villaverde Ramiro

Fdo: Aroa Palomo Vadillo

Firmado por PALOMO VADILLO AROA - ***3021** el día 04/07/2023 con un certificado emitido por AC FNMT Usuarios

Certificado del/los tutor/es TFG

D. José Manuel Sánchez Santos, profesor del Departamento de Estadística de la Universidad de Salamanca y Dª. Ángela Villaverde Ramiro investigadora del Centro de Investigación del Cáncer (CiC-IBMCC, USAL/CSIC),

HACE/N CONSTAR:

Que el trabajo titulado "*Análisis y caracterización de pacientes con leucemia mieloide aguda*", que se presenta, ha sido realizado por Aroa Palomo Vadillo, con DNI 70830217S, y constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Grado en Estadística en esta Universidad.

Salamanca, 3 de julio de 2023

Fdo.: José Manuel Sánchez Santos

Firmado digitalmente por SANCHEZ SANTOS JOSE MANUEL - 07870414K
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-07870414K, givenName=JOSE MANUEL, sn=SANCHEZ SANTOS, cn=SANCHEZ SANTOS JOSE MANUEL - 07870414K
Fecha: 2023.07.04 17:42:09 +02'00'

Fdo.: Ángela Villaverde Ramiro

Firmado por VILLAVERDE RAMIRO ANGELA - 06587720Z el día 04/07/2023 con un certificado emitido por AC FNMT Usuarios

ÍNDICE

RESUMEN

La leucemia mieloide aguda, AML, es un cáncer en las células productoras de la sangre que afecta generalmente a la sangre y a la medula ósea, siendo la leucemia aguda más común. Con unos datos de pacientes de esta enfermedad, recopilados de dos estudios previos, se ha querido estudiar mediante diferentes técnicas y análisis estadísticos los factores clínicos y genéticos más relevantes en el pronóstico y en la supervivencia de la AML. Así como llevar a cabo la validación de las categorías en las que se divide la AML según su orden de aparición, y los principales grupos de riesgo y las anomalías genéticas que los caracterizan, según la European LeukemiaNet 2017.

Palabras clave: Leucemia Mieloide Aguda, AML, European LeukemiaNet, pAML, sAML, Análisis de Supervivencia, Contrastes de Hipótesis, ETL, ISCN, Expresiones Regulares, RegExr.

ABSTRACT

Acute myeloid leukemia, AML, is a cancer of the blood-producing cells that generally affects the blood and bone marrow and is the most common acute leukemia. Using data from patients with this disease, collected from two previous studies, the aim was to use different techniques and statistical analyses to study the most relevant clinical and genetic factors in the prognosis and survival of AML. We also validated the categories into which AML is divided according to its order of appearance, and the main risk groups and the genetic abnormalities that characterize them, according to the European LeukemiaNet 2017.

Keywords: Acute Myeloid Leukemia, AML, European LeukeamiaNet, pAML, sAML, Survival Analysis, Hypotheses Testing, ETL, ISCN, Regular Expression, RegExr.

1. INTRODUCCIÓN

La leucemia es un tipo de cáncer de las células primitivas productoras de sangre, que afecta a las partes del sistema circulatorio e inmunológico del cuerpo, sangre y medula ósea principalmente. Se caracteriza por la producción descontrolada de células blancas anormales, especialmente leucocitos, los cuales son responsables de combatir infecciones y enfermedades. Se suele describir en función de su crecimiento clasificándola como aguda (de rápido crecimiento), donde se encuentran la leucemia linfocítica aguda y la leucemia mieloide aguda, y las leucemias crónicas (de lento crecimiento), dentro de las que están la leucemia linfocítica crónica y la leucemia mieloide crónica.

La leucemia mieloide aguda (AML) o leucemia mielocítica aguda es una neoplasia hematológica superpuesta, siendo la leucemia aguda más común en adultos y representando aproximadamente el 80% de los casos de este grupo. Es un trastorno clonal maligno caracterizado por alteraciones y baja producción de células madre hematopoyéticas saludables, que son las células inmaduras de la sangre que pueden transformarse en cualquier tipo de célula sanguínea (glóbulos blancos, glóbulos rojos y plaquetas). Debido a diversas alteraciones genéticas en los precursores de las células sanguíneas, las células hematopoyéticas saludables impiden la diferenciación de las células e impulsan la acumulación de blastos, reemplazando estos el tejido hematopoyético normal, lo que produce una sobreproducción de células madre mieloides clonales neoplásicas y la aparición de citopenias (bajo recuento de células sanguíneas). La acumulación de células inmaduras comienza en la médula ósea, provocando la supresión del crecimiento y diferenciación de las células sanguíneas normales, pero en la mayoría de los casos se acumula rápidamente en la sangre y a veces se propaga a diferentes partes del cuerpo como los ganglios linfáticos, bazo, hígado y el sistema nervioso central. (Prada-Arismendy et al., 2017)

Los síntomas resultantes consisten en diversos grados de anemia, neutropenia (bajo recuento de glóbulos blancos) y trombocitopenia (bajo recuento de plaquetas), así como la infiltración en los tejidos.

1.1. Diagnóstico, pronóstico y tratamiento.

Las leucemias mieloides agudas están asociadas con un inicio y una progresión rápidas y suelen presentar resistencia a la quimioterapia. La quimioterapia intensiva y tratamientos combinados han mostrado mejoras en la respuesta de los pacientes, aunque el riesgo de mortalidad relacionado con las recaídas sigue siendo muy alto.

La incidencia de la leucemia mieloide aguda (AML) es aproximadamente igual entre ambos sexos y aumenta en relación con la edad, siendo la edad promedio de los pacientes de en torno a 60 años. En pacientes menores de 65 años, se registra aproximadamente 1.3 caso por cada 100 000 habitantes, mientras que, en aquellos mayores de 65 años, la cifra asciende a 12.2 casos por cada 100 000 habitantes. A pesar de que los pacientes de edad avanzada predominan en esta enfermedad, están subrepresentados en la literatura debido a la selección de adultos más jóvenes para los ensayos clínicos.

Aunque se han logrado avances significativos en el tratamiento de la AML, lo cual ha mejorado los resultados para los pacientes más jóvenes, el pronóstico para los ancianos, que constituyen la mayoría de los nuevos casos, sigue siendo desfavorable. Incluso con los tratamientos actuales, se estima que hasta el 70% de los pacientes de 65 años o más fallecerán

debido a esta enfermedad un año después de su diagnóstico.(De Kouchkovsky & Abdul-Hay, 2016)

Los pacientes con AML se presentarán inicialmente de diversas maneras siendo algunos casos descubiertos mediante análisis de sangre rutinarios, mientras que otros pueden presentar complicaciones sintomáticas como infección, hemorragia o coagulación intravascular diseminada. Además, otros síntomas antecedentes serían fatigas e infiltraciones en la piel por la leucemia, conocida como leucemia cutis, ocurriendo en alrededor de la mitad de los pacientes con leucemias.

Actualmente, el diagnostico de AML se basa en una integración de métodos que permiten una caracterización exhaustiva y complementaria de cada caso, siendo estos la citomorfología, citoquímica, inmunofenotipado, citogenética y genética molecular. El diagnóstico se centra, en primer lugar, en el análisis de la médula ósea (MO) y la sangre periférica (SP) (recuento sanguíneo completo y recuento de blastos). El diagnóstico específico se confirma mediante esta integración de métodos, buscando la actividad de la mieloperoxidasa en los blastos, o mediante inmunofenotipado de moléculas de superficie como CD123, CD45, CD34, CD38, entre otros. (Prada-Arismendy et al., 2017)

El examen de la médula ósea desempeña un papel fundamental tanto para establecer el diagnóstico, como en la obtención de tejido para su análisis, lo que permite una mejor clasificación del subtipo de AML y determinar la gravedad del pronóstico.

Durante el diagnóstico, la clasificación y estratificación del riesgo es un momento crítico para la toma de decisiones sobre el tratamiento. Las decisiones sobre el tipo de quimioterapia, el momento del trasplante de células madre hematopoyéticas o la elección para ensayos clínicos se evalúan en función de la probabilidad, a priori, de cada paciente de lograr una remisión completa, la persistencia prospectiva de enfermedad residual medible y la probabilidad prevista de recaída o muerte. (Tazi et al., 2022)

La enfermedad subyacente en AML se debe a anormalidades en la producción celular hematológica, aunque pueden ocurrir manifestaciones extramedulares como sarcomas mieloides o leucemias cutis. Un pequeño subconjunto de casos tiene factores causantes identificados, como quimioterapia previa o exposición a ciertos productos químicos, sin embargo, la gran mayoría se debe a alteraciones genéticas, ya sea por anomalías cromosómicas o mutaciones genéticas aisladas, sin causa clara identificada. La delimitación de estas anomalías genéticas es importante para clasificar el riesgo de los pacientes y determinar el tratamiento adecuado. (Pelcovits & Niroula, 2010)

El objetivo del tratamiento consiste en lograr la remisión completa, es decir, erradicar las células neoplásicas y restaurar la hematopoyesis normal. Este depende de diversos factores, como la edad, el estado general de salud, las características citogenéticas y moleculares, y la presencia de mutaciones específicas. El tratamiento consta de varias etapas, siendo la primera la inducción, cuyo objetivo es reducir la carga de células leucémicas por debajo del nivel de detección. La remisión completa se define como la ausencia de células leucémicas en la médula ósea y la normalización de los recuentos de células sanguíneas periféricas. Sin embargo, generalmente se asume que existe una carga sustancial de células leucémicas que persiste sin ser detectada, lo que aumenta el riesgo de recaída en semanas o meses si no se administra más tratamiento.

Los fármacos estándar administrados en el tratamiento son daunorubicina (DNR) y citarabina (ara-C). La administración de DNR durante tres días junto con ara-C durante siete días en dosis "convencionales" ha logrado una remisión hematológica completa en aproximadamente el 70% de los pacientes menores de 60 años. La adición de otros agentes de quimioterapia a este tratamiento generalmente no ha mejorado la tasa o duración de la remisión

y en cambio, ha aumentado la toxicidad. Por tanto, esta combinación "3 + 7" de DNR y ara-C se ha convertido en el tratamiento estándar de inducción de remisión.

Una proporción significativa de pacientes que logran una remisión completa no puede recibir más terapia posterior debido a la toxicidad persistente del tratamiento inicial. Se debe tener en cuenta la agresividad de las toxicidades a la hora de aplicar cualquier tratamiento de cara a un tratamiento posterior efectivo, ya que la incapacidad para recibir este puede resultar en una recaída temprana.

Actualmente, se está probando en muchos centros la adición de nuevos fármacos en combinación de ara-C en dosis alta, como etopósido y ciclofosfamida o mitoxantrona con diaziquona. A pesar de obtener buenos resultados iniciales, la efectividad real de estos tratamientos aún está por determinarse, al igual que los subgrupos de pacientes que podrían beneficiarse más de un tratamiento que de otro.

Para la terapia posterior a la remisión se han seguido tres tipos de tratamiento:

- La terapia de consolidación o intensificación temprana. Consiste en la administración de quimioterapia igual, más intensa o no cruzada con la quimioterapia utilizada durante la fase de inducción. Cada ciclo de esta puede producir una mielosupresión profunda (anulación parcial o total de la actividad de la medula ósea).
- La quimioterapia de mantenimiento. Es un tratamiento que generalmente es menos mielosupresor, se administra con frecuencia durante varios meses o años, generalmente de forma ambulatoria.
- La quimioterapia mieloablativa intensiva o quimiorradioterapia, seguida de trasplante de médula ósea de un donante adecuado después de la inducción de la remisión.

La terapia citorreductora continua, trata de eliminar la mayor cantidad de tumor posible, desempeña un papel crucial en la prevención de la recaída de la AML. Numerosos estudios han demostrado que el tratamiento intensivo de consolidación es más efectivo que los tratamientos de mantenimiento o intensificación tardía. En ensayos no aleatorizados, se ha sugerido que los pacientes que reciben dosis altas de citarabina (ara-C) durante la terapia de consolidación experimentan remisiones más prolongadas en comparación con aquellos que reciben dosis estándar de ara-C.

El tratamiento posterior a la remisión varía según el grado de AML, el riesgo individual del paciente y factores específicos y será determinado por los oncólogos y hematólogos para cada caso particular. Existe controversia en cuanto a la estrategia de estos tratamientos más efectiva, ya que la mayoría de los investigadores y estudios sugieren que la terapia de mantenimiento no ha demostrado aumentar la tasa de curación en pacientes que ya han recibido quimioterapia intensiva de inducción y consolidación. Además, cabe destacar que la quimioterapia de inducción para la remisión de la AML no ha experimentado cambios significativos en más de una década. (Devine & Larson, 1994)

1.2. Biomarcadores.

La clasificación de las leucemias agudas, la estratificación del riesgo y el tratamiento de los pacientes se ha centrado tradicionalmente en características morfológicas y citoquímicas y marcadores citogenéticos. Sin embargo, gracias al desarrollo de nuevas tecnologías los avances en la fisiopatología de la AML a nivel celular, inmunología y biología molecular han mejorado nuestra comprensión de los marcadores biológicos que distinguen a las células hematopoyéticas normales de las leucémicas. Gracias a estas tecnologías la detección de otros marcadores

moleculares, como mutaciones puntuales y caracterización de perfiles epigenéticos y proteómicos, están teniendo gran relevancia y permiten la clasificación precisa de un clon maligno como mieloide, linfoide B, linfoide T o bifenotípico en la mayoría de los casos.

La identificación de estos nuevos biomarcadores contribuye a una mejor comprensión de las bases moleculares de la enfermedad, por lo que permitirá tomar decisiones óptimas de diagnóstico, monitoreo y tratamiento, mejorando así los resultados de los pacientes, así como predecir la respuesta al tratamiento de cada uno de ellos. Por ello las mutaciones genéticas están siendo incorporadas progresivamente en las pautas de clasificación y estratificación del riesgo en la AML. (Tazi et al., 2022)

Con el creciente uso de la secuenciación prospectiva en el diagnostico de AML, es importante entender la relevancia clínica de estos biomarcadores moleculares en relación con la MRD (enfermedad residual mínima), CR (remisión completa) y la recaída. Para llevar estos hallazgos a la práctica clínica se requiere desarrollar herramientas de apoyo a la toma de decisiones clínicas basadas en la evidencia y dinámica que consideren los biomarcadores moleculares y clínicos.

El diagnóstico de AML se establece cuando el 30% o más de todas las células nucleadas de la médula ósea son blastos. Los blastos de los pacientes con AML suelen ser más grandes que los linfoblastos y muestran una mayor heterogeneidad en cuanto a tamaño y forma, tienen un citoplasma más abundante y suelen contener gránulos citoplasmáticos. La sangre periférica suele contener células blastoides leucémicas en el 85-90% de los casos.

En relación con el recuento de glóbulos blancos, aproximadamente un tercio de los pacientes presenta niveles elevados, pero en los últimos años, los recuentos superiores a 100000/µl ocurren en menos del 10% de los casos, posiblemente debido a diagnósticos más tempranos. El recuento absoluto de granulocitos generalmente se encuentra deprimido en la AML, siendo inferior a 1,500/µl en la mitad de los pacientes al momento del diagnóstico.

Es común observar un grado moderado de anemia, y el recuento de plaquetas suele ser inferior a 100 000/µl, a menudo descendiendo incluso por debajo de los 20 000/µl. También se presentan elevaciones de leves a moderadas en los niveles séricos de ácido úrico, lo cual suele reflejar un aumento en la renovación celular.

La AML no se limita a la medula ósea y a la sangre periférica, ya que puede afectar a diferentes órganos debido a la infiltración de células leucémicas o complicaciones metabólicas relacionadas. La piel y los huesos son los lugares normalmente más afectados, aunque los sarcomas granulocíticos (tumores leucémicos que se forman fuera de la medula ósea y sangre) pueden ocurrir en cualquier órgano. La dificultad respiratoria en pacientes con AML se debe principalmente a infecciones, sin embargo, aquellos con un alto número de células blastoides circulantes (>100,000/µl) pueden desarrollar disnea grave e hipoxemia (nivel bajo de oxígeno en sangre) en los capilares pulmonares. (Devine & Larson, 1994)

El análisis citogenético ha ampliado nuestra comprensión de los procesos de transformación leucémica ya que, al estudiar minuciosamente los puntos de ruptura cromosómica, se ha logrado identificar genes que desempeñan un papel fundamental en ella. Se ha determinado la ubicación cromosómica de un gran número de estos oncogenes y aunque actualmente se conoce poco sobre su función precisa, se cree que muchos están involucrados en el control de la proliferación y diferenciación celular. Los cambios cromosómicos estructurales pueden activar o perturbar la expresión de oncogenes, lo que provoca alteraciones en la regulación celular y eventualmente en una transformación maligna.

Entre los oncogenes que más destacan en la AML están: (Prada-Arismendy et al., 2017)

- CEBPA. Las alteraciones genéticas relacionadas con estén gen están relacionadas con buen pronóstico de la enfermedad ya que es un gen supresor de tumores. Se encuentra aproximadamente en el 7-11% de los pacientes de AML. Los pacientes con doble mutación de este gen tienen riesgo bajo de presentar FLT3-ITD y es excluyente de NPM1.
- DNMT3A. La función de este gen está ligada a la renovación de células madre hematopoyéticas. Sus mutaciones son un evento temprano de AML y las más frecuentes después de NPM1 y FLT3 en la enfermedad. Este gen y sus mutaciones está relacionado con un efecto clínico negativo y una disminución de la supervivencia de los pacientes.
- IDH1/2. Las mutaciones de IDH se han detectado en el 15-20% de todos los pacientes con AML, la mayoría de cariotipo normal. Se asocian con un resultado clínico desfavorable, aunque dependen de las mutaciones de otros genes como NPM1 y FLT3 y el tipo.
- FLT3. Este gen regula la diferenciación y proliferación de las células progenitoras hematopoyéticas. La mutación más importante y frecuente encontrada en el pronóstico de AML es FLT3-ITD (aproximadamente 30-40% de los pacientes) aunque todas ellas están asociadas con mayor recuento de blastos, un mal pronóstico y menor supervivencia libre de enfermedad.
- NPM1. Funciona como supresor de tumores y sus alteraciones moleculares son frecuentes en pacientes con AML, sobre todo en aquellos con cariotipo normal (entre el 25-50%). Las mutaciones de este gen tienen impacto positivo en pacientes de AML si no están asociadas con FLT3-ITD, presentando remisión completa en el 85% de los casos y buena tasa de supervivencia libre de enfermedad.
- TET1/2. Las mutaciones de TET2 son mutuamente excluyente de IDH1/2. Se considera un evento temprano de la leucemogénesis en AML y está presente en el 10-20% de los pacientes. Se asocia estas mutaciones con una reducción de la supervivencia global de pacientes con riesgo favorable e intermedio y se correlaciona con mutaciones en NPM1.
- TP53. Este gen funciona como supresor de tumores y tiene uno de los factores pronósticos más comunes (75-78% de los casos) e importantes en la AML de cariotipo complejo (múltiples anormalidades citogenéticas). Las mutaciones son más comunes en la enfermedad relacionada con la terapia que en la inicial, y son un indicador de mal pronóstico. Los pacientes que la presentan suelen ser de mayor edad y con tasas de remisión y supervivencia bajas.

  RUNX. La familia de proteínas y mutaciones de este gen se conoce como factor de unión al núcleo (CBF) y lo codifican RUNX1, RUNX2, RUNX3. Más de 40 alteraciones cromosómicas involucran a RUNX1 como t(8;21) o la inv.(16)/t(16;16) y es característico de la AML-CBF (leucemia mieloide aguda del factor de unión al núcleo). Los pacientes con ella responden al tratamiento y obtienen remisión completa en un 80-90% de los casos, aunque casi la mitad de ellos experimentan recaídas.

Otros oncogenes y mutaciones importantes en AML son: ASXL1, CBL, GATA, KIT, NRAS.

1.3. Citogenética.

Las alteraciones cromosómicas son comunes en la AML y proporcionan información pronóstica y guían el tratamiento. Desde la implementación de las técnicas de bandeo cromosómico en la década de 1970, el análisis citogenético ha sido fundamental para la subclasificación clínica de las leucemias agudas. Se han identificado síndromes de leucemia aguda en los que anomalías cromosómicas específicas correlacionan con subtipos morfológicos y cuadros clínicos específicos. El papel más importante de la citogenética en la AML, aparte de su relevancia para la clasificación de la OMS, es la estratificación pronóstica, teniendo

relevancia también para el monitoreo de la cinética de la enfermedad, la evaluación de la respuesta y la caracterización de la evolución clonal. (Haferlach & Schmidts, 2020)

La citogenética se centra en el estudio de las anomalías cromosómicas en las células sanguíneas y engloba técnicas de análisis cromosómico e hibridación in situ con fluorescencia (FISH). El análisis cromosómico permite detectar células no malignas, las cuales suelen tener un cariotipo normal (46, XX o 46, XY), y cariotipos leucémicos, los cuales pueden mostrar alteraciones cromosómicas numéricas o estructurales adquiridas. El FISH utiliza sondas fluorescentes dirigidas contra locus cromosómicos específicos y pueden utilizarse para buscar aberraciones citogenéticas conocidas o sospechadas, así como aberraciones numéricas si se dirigen contra centrómeros. Esta técnica se puede realizar tanto en la interfase como en los cromosomas en metafase. Mientras que el análisis cromosómico permite una evaluación exhaustiva de todo el genoma, el FISH ofrece un enfoque más rápido y dirigido. El FISH de 24 colores permite caracterizar o validar aberraciones complejas encontradas en el análisis cromosómico después del bandeo.

En la mayoría de los casos de AML, se pueden detectar anomalías cromosómicas clonales que incluyen ganancias o pérdidas de cromosomas completos o de los brazos cortos (p) o largos (q) de los cromosomas, así como reordenamientos estructurales como translocaciones, inversiones o inserciones. Por ello, se recomienda la realización de análisis citogenéticos antes de iniciar el tratamiento en cada paciente recién diagnosticado, ya que estudios sobre el significado pronóstico de las anomalías citogenéticas recurrentes en AML han demostrado resultados similares. En muchos centros, los planes de terapia posterior a la remisión dependen de los resultados obtenidos en estos análisis al diagnóstico.

Se han identificado varias alteraciones cromosómicas recurrentes que tienen un peso pronóstico importante en cuanto a la respuesta a la quimioterapia y la supervivencia global entre las que se encuentran:

- t(6;9). Este síndrome se describe como una traslocación equilibrada entre los cromosomas 6 y 9 y está asociado a Leucemias Mieloides Agudas con basofilia (aumento de basófilos en sangre) de médula ósea. Representa a menos del 2% de los casos y no está asociada a ningún subgrupo particular, aunque suelen ser adultos jóvenes y que suelen mostrar mala respuesta a la quimioterapia.
- t(8;21)(q22;q22). Morfológicamente, la translocación t(8;21) está asociada con Leucemias Mieloides Agudas con Maduración, el tipo FAB M2, y ocurre en aproximadamente el 5% de los casos de AML afectando en mayor medida a adultos jóvenes. Este síndrome está asociado con el pronóstico más favorable teniendo los pacientes una alta tasa de remisión completa (85-100%) utilizando quimioterapia convencional.
- inv(16) o t(16;16). Este síndrome está asociado a la Leucemia Mielomonocítica Aguda con Eosinófilos Anormales (FAB M4Eo) e involucra anormalidades en ambos brazos del cromosoma 16, tanto la inversión inv(16)(p13q22) como la translocación t(16;16)(p13;q22). Los pacientes presentan una alta tasa de remisión completa y buen pronóstico.
- t(15;17). Este síndrome viene asociado a la Leucemia Promielocítica Aguda (FAB M3) e implica los cromosomas 15 y 17, produciendo coagulación intravascular diseminada al diagnóstico. Los pacientes no perteneces a un subgrupo particular y generalmente logran una alta tasa de remisión completa.
- t(9;11). Esta traslocación viene asociada a la leucemia Monoblástica Aguda (FAB M5) e involucra al brazo largo del cromosoma 11 (11q). La mayoría de los pacientes son niños o adultos jóvenes que suelen tener infiltraciones de células leucémicas a la piel y presentan remisiones, pero tienden a ser cortas. (Devine & Larson, 1994)

Las reordenaciones cromosómicas t(8;21) e inv(16) han sido reconocidas, según la clasificación de la OMS, como entidades distintas dentro de la categoría de 'AML con anomalías genéticas recurrentes'. Su presencia es específicamente diagnóstica de AML, incluso cuando los blastos de la médula ósea son inferiores al 20%. La AML con cualquiera de estas alteraciones citogenéticas también se ha incluido en el Grupo Genético Favorable según la clasificación de la Red Europea de Leucemia.

Los pacientes con CBF-AML suelen presentar recuentos más altos de glóbulos blancos y blastos en la médula ósea, y una probabilidad superior de tener enfermedad extramedular. Aunque un mayor número de pacientes con CBF-AML tienen más de 60 años, los pacientes con AML inv(16) suelen ser mayores que aquellos con t(8;21) en comparación. (Prada-Arismendy et al., 2017)

1.4. Categorías y fases. ELN

La clasificación tradicional de las leucemias agudas se ha basado en descripciones morfológicas, reflejando el tipo de célula predominante en la médula ósea. En 1976, un grupo de hematólogos formó el Grupo Cooperativo Francés-Americano-Británico (FAB) con el objetivo de establecer un sistema de subclasificación para las leucemias agudas que separara claramente la leucemia linfoblástica aguda (LLA) y la leucemia mieloide aguda (LMA) en trastornos distintos. (Devine & Larson, 1994)

Según el tipo de ocurrencia, esta enfermedad se clasifica en dos grandes grupos, AML primaria (pAML) y AML secundaria (sAML). La primaria se refiere a la leucemia mieloide aguda que surge de novo (o “de nuevo”) sin antecedentes y suele presentar mejor pronóstico que la secundaria. Esta, sin embargo, se refiere a un proceso leucémico que puede ser bien el resultado de una evolución de otro tipo de leucemia, con o sin tratamiento anterior, o bien ocurrir después de una previa exposición a radiación o quimioterapia para tratar otro tipo de cáncer. Debido a los peores resultados en la sAML, los tratamientos están cambiando actualmente y son diferentes a los utilizados en la pAML.

Aunque la presencia de alteraciones genéticas desempeña un papel crucial y, en parte, determinante para la clasificación de AML según la OMS, el diagnóstico se realiza mediante la presencia de ≥20% de blastos en la sangre periférica o en la médula ósea. Sin embargo, existen tres excepciones en las que el diagnóstico de AML se realiza independientemente del recuento de blastos (Haferlach & Schmidts, 2020):

- AML con t(8;21)(q22;q22.1); RUNX1-RUNX1T1.
- AML con inv(16)(p13.1q22) o t(16;16)(p13.1;q22); CBFB-MYH11.
- APL (leucemia promielocítica aguda) con t(15;17)(q22;q11-12); PML-RARA.

La Organización Mundial de la Salud en sus directrices actualizadas de 2016 distingue seis grupos de AML:

1. AML con alteraciones genéticas recurrentes.
2. AML con cambios relacionados con la mielodisplasia.
3. Neoplasias mieloides relacionadas con la terapia.
4. AML no especificada de otro modo.
5. Sarcoma mieloide.
6. Proliferaciones mieloides relacionadas con el síndrome de Down.

Los sistemas de estratificación de riesgo actualmente utilizados consideran las alteraciones citogenéticas y moleculares de alta relevancia pronóstica y tienen una gran capacidad predictiva en cuanto a tasas de remisión completa, supervivencia libre de

enfermedad, riesgo de recaída y supervivencia global. Los dos modelos más importantes actualmente son la estratificación de riesgo recomendada por la European LeukemiaNet (ELN) (Döhner et al., 2017) y el modelo de riesgo del Medical Research Council (MRC) (Grimwade et al., 2016). Estos sistemas siguen siendo demasiado simples ya que, aunque incorporan alteraciones citogenéticas y moleculares no incorporan parámetros clínicos ni específicos del paciente. (Haferlach & Schmidts, 2020)

Las pautas clínicas actuales en el AML reconocen tres grupos de riesgo citogenético: favorable, intermedio y desfavorable.

Tabla 1. Clasificación de la ELN 2017

Grupo de riesgo	Anomalía genética
Favorable	t(8;21)(q22;q22.1); RUNX1-RUNX1T1
	inv(16)(p13.1q22) ó t(16;16)(p13.1;q22); CBFB-MYH11
	Mutación de NPM1 en ausencia de FLT3-ITD o bajo FLT3-ITD
	Mutación de CEBPA bialélica aislada
Intermedio	Mutación de NPM1 y FLT3-ITD (alta)
	NPM1 tipo salvaje sin FLT3-ITD ó bajo FLT3-ITD (cariotipo normal)
	t(9;11)(p21.3;q23.3); MLLT3-KMT2A
	Anomalías genéticas no clasificadas como favorable o desfavorable
Desfavorable	t(6;9)(p23;q34.1); DEK-NUP214
	t(v;11q23.3); KMT2A reorganizado
	t(9;22)(q34.1;q11.2); BCR-ABL1
	inv(3)(q21.3q26.2) ó t(3;3)(q21.3;q26.2); GATA2,MECOM(EVI1)
	-5 ó del(5q); -7; -17/abn(17p)
	Cariotipo complejo
	Cariotipo con monosomías
	NPM1 tipo salvaje y alto FLT3-ITD
	Mutación de RUNX1
	Mutación de ASXL1
	Mutación de TP53

Nota. Clasificación de las principales anomalías genéticas y biomarcadores para la estratificación del riesgo según la ELN 2017.

1.5. Objetivos

El objetivo principal del presente trabajo consiste, mediante técnicas estadísticas con una cohorte de pacientes, en la validación de las categorías de riesgo actuales que clasifican a los pacientes de leucemia mieloide aguda según su supervivencia por la European LeukemiaNet (ELN) 2017. Así como confirmar las anomalías citogenéticas clasificadas dentro de cada uno de los grupos de estratificación del riesgo.

A partir de este objetivo, surge como secundario el de encontrar nuevos patrones o factores, clínicos o genéticos, tanto de riesgo como favorables en el curso y pronóstico de la enfermedad.

Estos objetivos se llevarán a cabo con el fin de corroborar, con unos datos de pacientes con AML de distinta procedencia, las conclusiones que aparecen en la teoría y últimos estudios sobre la enfermedad, al igual que, si se diera el caso elaborar nuevas afirmaciones sobre la supervivencia y factores perjudiciales o beneficiosos en esta.

2. MATERIALES Y MÉTODOS

2.1. Base de datos

La base de datos utilizada en este estudio se ha obtenido como resultado de combinar dos bases de datos de dos estudios previos.

El primer estudio se basa en la utilización de métodos de aprendizaje automático para integrar firmas genómicas y así mejorar la clasificación y subclasificación de leucemia mieloide aguda más allá de las categorías primaria y secundaria. Mediante algoritmos de aprendizaje automático se analizaron firmas genómicas para identificar patrones distintivos y características únicas de los datos y utilizarlos para crear subgrupos más precisos y refinados dentro de las categorías ya conocidas. La integración de firmas genómicas permitió obtener una mejor precisión en la subclasificación y proporcionó información adicional sobre las características genéticas y biológicas de cada subgrupo. Los datos recogidos en él son una combinación de datos de pacientes provenientes del área oncológica AML del hospital Cleveland Clinic, del laboratorio Munich Leukemia Laboratory y datos públicos (The Cancer Genome Atlas, el ensayo maestro BEAT AML y el Grupo de Estudio Alemán-Austríaco) obteniendo el total de 6788 pacientes. En él se hizo un seguimiento de los pacientes durante una media de 12,4 meses, recogiendo periódicamente muestras de sangre periférica y/o médula ósea y excluyendo los casos que no tenían recogida la citogenética del paciente. (Awada et al., 2021)

El segundo estudio se basa en la clasificación y estratificación del riesgo unificado en AML, siendo un enfoque para categorizar y evaluar el riesgo en los pacientes, teniendo como objetivo agrupar a los pacientes en diferentes categorías que reflejen sus características genéticas, moleculares y clínicas y permitan así, predecir el pronóstico y guiar el tratamiento. La clasificación y estratificación del riesgo proporciona una base sólida para la toma de decisiones clínicas, predicción del pronóstico y personalización de tratamientos, con el fin de mejorar la compresión de la enfermedad y del manejo clínico para unos mejores resultados terapéuticos. Los datos recogidos en él incluyeron a 2113 paciente adultos con la enfermedad inscritos en ensayos del UK National Cancer Research Institute, el cual recoge hasta el 80% de pacientes de Reino Unido aptos para recibir tratamiento intensivo o no intensivo, por lo que son representativos de la población ya que no son limitados por criterios de ingreso a ensayos clínicos. La evaluación de la cohorte incluyó cariotipos, alteraciones en el nuero de copias y mutaciones oncogénicas en el cuerpo de diversos genes implicados en patogénesis de neoplasias mieloides en el diagnostico. (Tazi et al., 2022)

La base de datos resultante como unión de ambas se llevó a cabo mediante un proceso ETL, obteniendo de la primera base en concreto el factor citogenético (fórmula del cariotipo), de la segunda los factores clínicos (ej. Edad, género, recuento de plaquetas, glóbulos blancos, etc.) y uniendo en común los genes, anomalías del cariotipo, tipo de leucemia, etc.

Listado de variables de la base final:

ID: identificador del paciente.
Age: variable numérica que indica la edad del paciente en años.
Gender: variable categórica que indica el género del paciente siendo el valor 1=hombre y 0=mujer.
Os_year: variable numérica, contiene la supervivencia del paciente en años.
Os_status: variable de interés, indica si el paciente ha fallecido (=1) o sigue vivo (=0).
AML_Type: variable categórica que indica la clasificación de la leucemia en pAML (primaria) o sAML (Secundaria, deriva de otra, otro tipo de cáncer o es metástasis).

Wbc: variable numérica que indica los glóbulos blancos en 10^9/L.
Hb: variable numérica que indica la hemoglobina en g/dL.
Plt: variable numérica que indica las plaquetas en 10^9/L .
Bm_blasts: variable numérica que indica el porcentaje de blastos.
Eln_2017: variable categórica que indica mediante el European LeukemiaNet (ELN) la clasificación del estado del paciente en adverse/intermediate/favorable según las diferentes anomalías genéticas que tiene el paciente.
Cytogenetics: fórmula del cariotipo bajo la nomenclatura ISCN.

Anomalías genéticas del cariotipo clasificadas en variables categóricas siendo los valores 1=tenerlo y 0=no tenerlo:

Complex, t(6;9), minus_5_del(5), minus_7_del(7), minus_9_del(9), del(12), del(13), del(16), minus_17_del(17), inv(3)_t(3;3), minus_y

Genes:

ASXL1, CBL, Bi-CEBPA, M-CEBPA, DNMT3A, ETV6, EZH2, FLT3, GATA2, IDH1, IDH2, KIT, KRAS, NPM1, NRAS, PHF6, PRPF8, PTPN11, RAD21, RUNX1, SETBP1, SF3B1, SMC3, SRSF2, STAG2, TET2, TP53, U2AF1, WT1, ZRSR2

2.2. Proceso ETL

Los procesos ETL (Extract - Transform - Load), extracción, transformación y carga, son utilizados en la integración y creación de almacenes de datos. Los datos son extraídos de diversas fuentes, transformados para cumplir ciertos requisitos y cargados en un sistema de destino, como almacenes o bases. (El-Sappagh et al., 2011)

1. Extracción (E)

   Consiste en la identificación de las fuentes de donde provienen los datos, como bases, archivos, APIs, etc., y establecer conexiones con estos para así recuperar los datos requeridos. Una vez obtenidos de las fuentes utilizando diversas técnicas, como consultas o extracción de archivos, se capturan los datos necesarios de origen.

2. Transformación (T)

   Los datos extraídos son limpiados para eliminar cualquier error, inconsistencia, duplicado o información irrelevante y así garantizar la calidad y consistencia de estos. Tras ello, los datos son integrados, ya que al ser una combinación de diferentes fuentes se requiere estandarizar formatos y fusionarlos.

   Para hacer más enriquecidos los datos se les puede agregar transformaciones o cálculos. Por último, los datos son validados según reglas o restricciones predefinidas para garantizar su precisión e integridad, y si fuera necesario, anonimizar o enmascarar datos sensibles para proteger la privacidad o cumplir regulaciones.

3. Carga (L)

   Se define la estructura y formato del sistema de destino, ya sean almacenes de datos o bases, donde se cargarán los datos ya transformados. Estos son mapeados en el sistema de destino para que coincidan adecuadamente los atributos y relaciones y luego se cargan, ya sea por lotes o en tiempo real.

   Tras esto se realizan comprobaciones de calidad de los datos verificando la integridad, precisión y coherencia dentro del sistema de destino de los datos. Por último, se aplican técnicas de indexación u optimización de rendimiento para facilitar la consulta y análisis posterior de los datos cargados.

Estos procesos a menudo son automatizados para manejar grandes volúmenes de datos de manera rápida y eficiente y programar actualizaciones regulares de los datos.

ETL es un parte fundamental de la integración y creación de almacenes de datos, ya que permite consolidar y analizar datos de diversos orígenes y obtener información relevante en los procesos de toma de decisiones.

2.3. Anomalías del cariotipo

El cariotipo es la disposición organizada de los cromosomas en una célula, generalmente representado visualmente.

Fórmula ISCN

Una nomenclatura estandarizada es fundamental para la descripción precisa y coherente de los cambios genómicos identificados mediante cariotipado, hibridación fluorescente in situ y microarrays, así como clasificar las diferentes anomalías en los cromosomas y cariotipos humanos. El Sistema Internacional de Nomenclatura Citogenómica Humana (ISCN) es la referencia central para la descripción del cromosoma humano, resultados de cariotipado, FISH y microarrays desde 1960. Proporciona de una manera consistente y reconocida internacionalmente, reglas para describir los hallazgos citogenéticos y citogenéticos moleculares en informes de laboratorio. Estos informes son documentos para el clínico remitente, y deben ser claros, precisos y contener toda la información relevante para una buena interpretación de los hallazgos citogenéticos.

Este sistema ha evolucionado en múltiples ocasiones ya que tiene que ir actualizándose a medida que el campo de la citogenética sigue ampliándose con nuevas tecnologías de base molecular. Por ello las revisiones de la nomenclatura citogenética deben ser cada vez más concurrentes y exhaustivas. Además, el uso de la citogenética en oncología continúa creciendo por lo que la nomenclatura citogenética para las neoplasias debe continuar estando bien definida.

El comité de la ISCN está compuesto por expertos (cito)genetistas que representan a cada continente y se reúne periódicamente para abordar cambios en el campo, que se resumen en versiones actualizadas de la ISCN. (Simons et al., 2013)

La fórmula ISCN para describir el cariotipo consta de varios componentes como el número de cromosomas, la constitución de los cromosomas sexuales, anomalías estructurales (ej. deleciones, duplicaciones, traslocaciones, etc.) y anomalías numéricas (ej. monosomías, trisomías, etc.).

El ISCN 2016 se actualizó respecto de los anteriores sistemas proporcionando pautas para la descripción y el informe de diversas anomalías citogenéticas específicas, incluyendo cambios numéricos y estructurales en los cromosomas. Incluye reglas de nomenclatura, patrones de bandeo e interpretación de resultados. (McGowan-Jordan et al., 2016)

Un ejemplo de fórmula del cariotipo bajo la nomenclatura ISCN 2016 podía ser: 46,XX,t(3;3)(q21;q26)[21], representando a una mujer (46, XX) con una translocación entre el brazo largo del cromosoma 3(q21) y el brazo largo del cromosoma 3(q26) y un numero de células de 21.

El sistema ISCN permite también hacer descripciones más complejas con detalles adicionales como implicaciones de otros cromosomas, puntos de rotura entre ellos, regiones

específicas del cromosoma afectadas, etc. Por lo que la fórmula puede variar según presente o no estas anomalías específicas.

2.4. Expresiones regulares

Las expresiones regulares, conocidas como “regex” o “regexp”, son una herramienta potente, flexible y eficiente de búsqueda, extracción y manipulación de cadenas de texto basadas en patrones específicos o reglas. Son usadas en varios lenguajes de programación, editores de textos y herramientas en línea.

En programación representan secuencias de caracteres que describen conjuntos de cadenas sin enumerar sus elementos y haciendo uso de la sintaxis propia de los diferentes lenguajes de programación. Su utilización requiere conocimiento de algún lenguaje de programación sencillo para poder llevar a cabo búsquedas de textos determinados mediante herramientas informáticas, para poder llevar a cabo tareas como adición, aislamiento o supresión de textos de manera rápida y ágil. (Chacón Beltrán, 2008)

Los patrones de las expresiones pueden incluir caracteres literales o normales de texto, metacaracteres y secuencias especiales. Los metacaracteres son caracteres con significado especial, como asteriscos (*), interrogaciones (?), mientras que las secuencias especiales son patrones predefinidos como \d (cualquier dígito) o \s (cualquier carácter de espacio en blanco).

Las principales operaciones que se pueden realizar como se ha mencionado previamente son las siguientes:

- Coincidencia: determinar si una cadena de texto cumple un patrón específico.
- Búsqueda y reemplazo: búsqueda de patrones específicos en un texto y reemplazo de su contenido o extracción de datos específicos.
- Validación y extracción de datos: validar la entrada de usuarios, como códigos postales, y extraer las partes específicas, como grupos entre paréntesis.
- Tokenización: las expresiones regulares pueden dividir una cadena de texto en cadenas más pequeñas basados en patrones específicos. Útil para el análisis de procesamiento de datos estructurados en texto. (Friedl, 2006)

En este caso en específico, se ha usado para mediante la búsqueda y reemplazo llevar a cabo correcciones de la formula del cariotipo bajo la nomenclatura ISCN 2016, para posteriormente mediante coincidencia de patrones buscar y clasificar las diferentes anomalías de esta en columnas en una tabla.

Como ejemplo tenemos la fórmula “46,XX,t(3;3)(q21;q26)[21]“ la cual corrige y clasifica en columnas los clones, numero de células, numero de cromosomas, cromosoma sexual, y la traslocación dividiéndola también en cromosoma y banda.

Ilustración 1. Corrección de la fórmula ISCN

orig	iscn	clones	ncells	anomaly	nc	sex	translocation	chromosome	band
46,XX,t(3;3)(q21;q26)[21]	46,xx,t(3;3)(q21;q26),[21]	46,xx,t(3;3)(q21;q26),[21]	[21]	46	46				
46,XX,t(3;3)(q21;q26)[21]	46,xx,t(3;3)(q21;q26),[21]	46,xx,t(3;3)(q21;q26),[21]	[21]	xx		xx			
46,XX,t(3;3)(q21;q26)[21]	46,xx,t(3;3)(q21;q26),[21]	46,xx,t(3;3)(q21;q26),[21]	[21]	t(3;3)(q21;q26)			t(3;3)(q21;q26)	3;3	q21;q26
46,XX,t(3;3)(q21;q26)[21]	46,xx,t(3;3)(q21;q26),[21]	46,xx,t(3;3)(q21;q26),[21]	[21]	[21]					

Nota. Ejemplo de corrección de la fórmula del cariotipo bajo la nomenclatura ISCN2016 con expresiones regulares. Elaboración propia.

2.5. Contrastes de hipótesis

Para analizar las posibles relaciones entre las variables y poder sacar conclusiones y continuar con otros análisis posteriores, se han efectuado diferentes pruebas estadísticas según la naturaleza de las variables a comparar.

VARIABLES NUMÉRICAS

Previamente, se comprobará la normalidad y la homocedasticidad de las variables para poder elegir correctamente los métodos paramétricos o no paramétricos.

2.5.1. Contraste de normalidad. Test de Shapiro-Wilk.

La distribución normal o Gaussiana es un tipo de distribución de probabilidad de variables continuas y está determinada por dos parámetros, la media y la desviación típica. Su función de densidad viene dada por la ecuación:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \; para\ todo\ x \in \mathbb{R} \tag{2.1}$$

por lo que si una variable, X sigue una distribución normal de media μ y varianza σ^2, se expresa como $X \approx N(\mu, \sigma)$.

Para contrastar si una muestra $\{X_1, \dots, X_n\}$ de tamaño n de una variable numérica proviene de esta distribución se proponen las siguientes hipótesis:

$$\begin{cases} H_0: X = N(\mu, \sigma) \\ H_1: X \neq N(\mu, \sigma) \end{cases} \tag{2.2}$$

El test de Shapiro-Wilk fue desarrollado en 1965 por los estadísticos Samuel Shapiro y Martin Wilk (Shapiro & Wilk, 1965) y es utilizado en diversos campos para comprobar la normalidad de un conjunto de datos, siendo considerado uno de los más potentes y precisos.

Considerando los datos estandarizados mediante la media y la varianza muestrales, $\bar{X}$ y S^2:

$$Z_i = \frac{X_i - \bar{X}}{S}, i \in 1, \dots, n \tag{2.3}$$

Se construye el estadístico W:

$$W = \frac{\left(\sum_{i=1}^{k} a_{n-i+1} \cdot (x_{n-i+1} - x_i)\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.4}$$

$$\text{donde k} = \frac{\text{n}}{2} \text{ si n es par y k} = \frac{\text{n} - 1}{2} \text{ si n es impar}$$

siendo los coeficientes a_i calculados como:

$$a = (a_1, a_2, \dots, a_n) = \frac{m^2 V^{-1}}{\|V^{-1} m\|} \quad (2.5)$$

donde m es el vector de medias, $m = (m_1, m_2, \dots, m_n)^T$

y V la matriz de covarianzas, $V = cov(z_i)$

El estadístico W verifica que $0 \leq W \leq 1$, siendo 1 si la muestra cumple que es normal, así que rechazamos los valores cercanos a 0 para aceptar la normalidad, es decir, es un contraste unilateral inferior, por tanto, comparando el valor observado en la muestra W_{obs} con el valor crítico $W_{n,\alpha}$, si $W_{obs} > W_{n,\alpha}$ no se rechazará la hipótesis nula por lo que provendrá de una muestra normal.

La función de R utilizada para su cálculo será: *shapiro.test(x)*.

2.5.2. Análisis de las varianzas. Test de Bartlett y test de Levene.

Se contrastará si las varianzas de las variables numéricas a estudiar pueden considerarse iguales (homocedasticidad) o diferentes (heterocedasticidad) entre las k categorías de una variable cualitativa. Esta suposición es importante para poder obtener inferencias estadísticas robustas posteriormente, ya que si no se obtendrían resultados sesgados e incorrectos. Para realizar este contraste se elaboran las siguientes hipótesis:

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \\ H_1: \sigma_i^2 \neq \sigma_j^2 \; para \; algún \; i \neq j, i, j = 1,2, \dots, k \end{cases} \quad (2.6)$$

2.5.2.1. Test de Bartlett

El test de Bartlett surge en 1937 por el estadístico Maurice S. Bartlett, es una de las pruebas más utilizadas para probar la igualdad de varianzas de k muestras entre grupos, en contraposición de que estas son desiguales en al menos dos de ellos (Bartlett, 1947).

Fue propuesto para probar la homocedasticidad de dos poblaciones que seguían una distribución normal, aunque posteriormente Snedecor y Cochran (1989) lo extendieron para múltiples poblaciones. (Odoi et al., 2022). Este test no requiere que los tamaños de muestra de cada grupo sean iguales, pero ninguno de ellos debe ser inferior a 3 y la mayoría debe ser superior a 5.

Esta prueba depende de la curtosis de las distribuciones y es más sensible a la falta de normalidad, por lo que cuando las distribuciones de las poblaciones se alejan de la normalidad tiene resultados deficientes, por ello en este estudio se utilizará cuando previamente se haya probado el supuesto de normalidad de los datos. (Jayalath et al., 2017)

Viene dado por el estadístico:

$$B = -\frac{\sum_{i=1}^{k}(n_i - 1)\ln(s_i^2/s_p^2)}{1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}(\frac{1}{n_i - 1}) - \frac{1}{\sum_{i=1}^{k}(n_i - 1)}\right]}, \quad (2.7)$$

donde s_i^2 es la varianza muestral del i-ésimo grupo:

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 \qquad (2.8)$$

y la varianza agrupada s_p^2 es el promedio ponderado de las varianzas muestrales:

$$s_p^2 = \sum_{i=1}^{k}(n_i - 1)s_i^2 \,/\, \sum_{i=1}^{k}(n_i - 1). \qquad (2.9)$$

Las varianzas serán significativamente diferentes al nivel de significación α si $B > \mathcal{X}^2_{(\alpha,\alpha-1)}$, donde $\mathcal{X}^2_{(p,k)}$ es el p-ésimo percentil superior de la distribución chi-cuadrado con k grados de libertad.

La función de R utilizada para su cálculo será: *bartlett.test(x, y)*.

2.5.2.2. Test de Levene

El test de Levene surge en 1960 por Howard Levene y se ha ido modificando a lo largo de los años por Brown y Forsythe (1974), Lim y Loh (1996), Hines (2000) y Parra-Frutos (2009) hasta lograr que sea una prueba robusta ante el supuesto de no normalidad o tamaños de muestra desiguales. (Jayalath et al., 2017)

Esta prueba deriva de la distribución *F* en el análisis de la varianza de un factor entre grupos (ANOVA), donde la observación x_{ij} se reemplaza por la desviación absoluta de la media de grupos: $z_{ij} = |x_{ij} - \bar{x}_i|$ siendo $\bar{x}_i$ el promedio muestral del i-ésimo grupo.

El estadístico de Levene viene dado por la fórmula:

$$W = \frac{N-k}{k-1} \cdot \frac{\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{k}(Z_{ij} - Z_{i.})^2} \qquad (2.10)$$

donde *k* es el número de grupos diferentes, N_i el numero de casos en el i-ésimo grupo, *N* el número total de casos en todos los grupos y Y_{ij} el valor de la variable medida en el j-ésimo caso del i-ésimo grupo.

Las varianzas serán significativamente diferentes al nivel de significación α elegido como resultado de la aproximación de *W* a la distribución *F* con *k-1* y *N-k* grados de libertad, $F_{(1-\alpha;\, k-1, N-k)}$.

La función de R utilizada para su cálculo será: *leveneTest (x ~ y)*.

2.5.3. Prueba no paramétrica. Prueba U de Mann-Whitney

La prueba U de Mann-Whitney o también conocida como "Test de Mann-Whitney-Wilcoxon" (WMW) o "Wilcoxon Rank Sum Test" surge como la alternativa no paramétrica a la prueba t de Student para muestras independientes ya que asume que las muestras no tienen por qué seguir una distribución conocida. Fue propuesta en 1945 por Frank Wilcoxon para el caso de igual tamaño de muestras y posteriormente ampliada en 1947 por Henry Mann y Donald Ransom Whitney para diversos tamaños de muestra. Se utiliza principalmente para comprobar si existen diferencias entre las distribuciones de dos muestras X e Y independientes o contrastar las medianas (θ) de dos muestras independientes. (Hettmansperger & McKean, 1998)

Siendo las hipótesis planteadas como:

$$\begin{cases} H_0: P(X > Y) = P(Y > X) \\ H_1: P(X > Y) \neq P(Y > X) \end{cases} \qquad \begin{cases} H_0: \theta_X = \theta_Y \\ H_1: \theta_X \neq \theta_Y \end{cases} \tag{2.11}$$

Para el cálculo del estadístico U es necesario previamente juntar, ordenar de menor a mayor y asignar el rango que ocupan las n observaciones, mezclando los $(X_1, X_2, \dots, X_{n_1})$ con los $(Y_1, Y_2, \dots, Y_{n_2})$, siendo los tamaños muestrales n_1 y n_2 respectivamente, por lo que $n = n_1 + n_2$. De encontrar valores repetidos, conocidos como empates o "ties", se les asignará como rango el rango medio entre las posiciones que se encuentre.

El estadístico U viene dado por la fórmula:

$$U = \min(U_1 + U_2) \tag{2.12}$$

siendo

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} \tag{2.13}$$

$$U_2 = R_2 - \frac{n_2(n_2 + 1)}{2} \tag{2.14}$$

donde R_1 = *suma de los rangos de las observaciones de X* y R_2 = *suma de los rangos de las observaciones de Y.*

Para la obtención del p-valor y contrastar las hipótesis al nivel de significancia α elegido, siendo $m = \text{máx}\{n_1, n_2\}\, y\ \ n = \text{mín}\{n_1, n_2\}$ se calculan las significaciones inferiores y superiores,

$$P_I = P(U \leq U_{obs}) \qquad P_S = 1 - P(U \leq U_{obs} - 1)$$

para el contraste bilateral se compara la probabilidad $P = 2 \cdot \min\{P_I, P_S\}$ con los valores de la tabla U de Mann-Whitney, rechazando la hipótesis nula cuando $U_{obs} \geq U_{crit}$.

La función en R utilizada para su cálculo será: *wilcox.test (x, y, alternative).*

2.5.4. Prueba paramétrica. T test

El T-test y la distribución T de Student fueron desarrollados por el estadístico William Sealy Gosset en 1908, presentado en el artículo "*The Probable Error of a Mean*" (Student, 1908) y posteriormente refinado por Ronald Fisher (Fisher, 1925). Surgió como solución al problema de utilizar la distribución normal en muestras pequeñas y con varianza poblacional desconocida. En la distribución T de Student los datos siguen una distribución normal y al poder trabajar con muestras pequeñas permite hacer inferencias precisas sobre las medias de la población. Se utiliza principalmente para comparar medias entre dos grupos, ya sean independientes o relacionados, ya que permite determinar si la diferencia observada entre ellas es significativa o debida al azar (Livingston, 2004) (Kim, 2015). En este estudio será utilizado el test para muestras independientes.

Dadas dos muestras independientes X e Y, que provienen de dos poblaciones con medias μ_X, μ_Y y varianzas σ_1^2, σ_2^2 respectivamente, las hipótesis planteadas para esta prueba son:

$$\begin{cases} H_0: \mu_X = \mu_Y \\ H_1: \mu_X \neq \mu_Y \end{cases} \tag{2.15}$$

El estadístico *T* viene dado por:

- cuando las varianzas son desconocidas pero iguales ($\sigma_1^2 = \sigma_2^2$)

$$T = \frac{\overline{X_1} - \overline{X_2}}{Sp \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2.16)$$

$$con\ \ Sp = \sqrt{\frac{(n_1 - 1)S_{C1}^2 + (n_2 - 1)S_{C2}^2}{n_1 + n_2 - 2}} \quad (2.17)$$

- cuando las varianzas son desconocidas pero distintas ($\sigma_1^2 \neq \sigma_2^2$)

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_{C1}^2}{n_1} + \frac{S_{C2}^2}{n_2}}} \quad (2.18)$$

donde $\overline{X_1}, \overline{X_2}$ son las medias muestrales, S_{C1}^2, S_{C2}^2 son las cuasivarianzas y n_1, n_2 los tamaños muestrales de X e Y respectivamente.

Para contrastar las hipótesis al nivel de significancia α elegido se realiza una prueba bilateral T con $n_1 + n_2 - 2$ grados de libertad, rechazando la hipótesis nula cuando el estadístico T_{obs} se encuentra fuera de la región de aceptación $(-t_{n_1+n_2-2,1-\alpha}, t_{n_1+n_2-2,1-\alpha})$.

La función en R utilizada para su cálculo será: *t.test (x , y)*

VARIABLES CATEGÓRICAS

2.5.5. Prueba para dos categorías. Chi-cuadrado.

La prueba chi-cuadrado, χ^2, fue desarrollada por Karl Pearson en 1900 y es una prueba no paramétrica para determinar si existe asociación entre variables nominales que se distribuye según una Chi-cuadrado conocida como uno de los "estadísticos de Pearson". Los datos suelen ordenarse en una tabla de contingencia y la prueba trata de evaluar si entre las frecuencias observadas y la esperadas existen diferencias significativas bajo una hipótesis nula. Esta prueba es sensible a los tamaños de muestra, resultando mejor en tamaños más grandes, y no requiere igualdad de varianzas entre los grupos de estudio ni homocedasticidad de los datos. Entre sus ventajas destaca su robustez respecto a la distribución que siguen los datos, su uso en estudios en los que no se cumple el supuesto paramétrico, la facilidad de cálculo y que proporciona información detallada sobre qué categorías exactamente son responsables de las diferencias encontradas. (McHugh, 2013)

Siendo las hipótesis planteadas como:

$$\begin{cases} H_0: \forall x, y\ \ f(x,y) = f(x) \cdot f(y) & X\ e\ Y\ son\ independientes \\ H_1: \forall x, y\ \ f(x,y) \neq f(x) \cdot f(y) & X\ e\ Y\ no\ son\ independientes \end{cases} \quad (2.19)$$

Se elabora una tabla de contingencia con las frecuencias observadas de cada grupo $X_i\ e\ Y_j$ siendo estas $O_{ij} = n_{ij}$. Se calculan las frecuencias esperadas estimadas bajo el supuesto de independencia

$$\widehat{E_{\iota}} = n \cdot \hat{p}_{ij} = \frac{n_i \cdot n_j}{n}. \quad (2.20)$$

El estadístico de Pearson viene dado por:

$$\chi^2_{obs} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2.21)$$

este se distribuye siguiendo una distribución $\chi^2_{(r-1)\cdot(s-1)}$ donde r y s son las probabilidades $\hat{p}_i$ y $\hat{p}_j$ respectivamente.

El estadístico será significativo al nivel de significancia teórica α elegido si resolviendo el contraste con las tablas chi-cuadrado $\chi^2_{obs} \geq \chi^2_{(r-1)\cdot(s-1);1-\alpha}$ y, por tanto, X e Y independientes.

La función en R utilizada para su cálculo será: *chisq.test (x)*.

2.5.6. Prueba no paramétrica para k categorías independientes. Prueba de Kruskal-Wallis.

El test de Kruskal-Wallis fue desarrollado por William Kruskal y W. Allen Wallis en 1952 y surge como una extensión de la prueba U de Mann-Whitney para más de dos muestras independientes. Esta prueba es el equivalente no paramétrico del test de análisis de la varianza, ANOVA, el cual supone la normalidad de los datos y homogeneidad de las varianzas. Sin embargo, el test de Kruskal-Wallis maneja datos que no cumplen estas suposiciones, aunque asume que las observaciones de cada grupo provienen de poblaciones con la misma distribución y que las muestras son aleatorias e independientes. Su finalidad es probar si las muestras provienen de la misma distribución, así como sus comparar medianas. (McKight & Najab, 2010)

Siendo las hipótesis de la prueba planteadas como:

$$\begin{cases} H_0: F_1(x) = F_2(x) = \cdots = F_k(x) \\ H_1: F_1(x) \neq F_2(x) \neq \cdots \neq F_k(x) \end{cases} \quad (2.22) \qquad \begin{cases} H_0: \theta_1 = \theta_2 = \cdots = \theta_k \\ H_1: \theta_1 \neq \theta_2 \neq \cdots \neq \theta_k \end{cases} \quad (2.23)$$

Tras ordenar todos los datos de menor a mayor y calcular sus rangos, siendo R_{ij} el rango del dato R_{ij}, el estadístico de Kruskal-Wallis, *H*, viene dado por:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j (\bar{R}_{.j} - \bar{R}_{..})^2 = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j (\bar{R}_{.j} - \frac{N+1}{2})^2 \quad (2.24)$$

donde $R_{.j}$ es la suma de los rangos de la muestra j y $\bar{R}_{.j}$ el rango medio de la muestra j, y $\bar{R}_{..}$el rango medio global.

El estadístico será significativo al nivel de significancia teórica α elegido si resolviendo el contraste el estadístico tiene valores grandes y alejados de 0 y buscando en la tabla H el valor crítico se cumple que $H_{obs} \geq H_{crit}$, por tanto, se rechazará la hipótesis nula.

También se cumple que el estadístico de Kruskal-Wallis sigue una distribución chi-cuadrado con k-1 grados de libertad, $H \approx \chi^2_{k-1;1-\alpha}$. Se rechazará la hipótesis nula cuando $H_{obs} \geq \chi^2_{k-1;1-\alpha}$.

La función en R utilizada para su cálculo y para el post-hoc serán respectivamente: *kruskal.test (x, g ,...), pairwise.wilcox.test(x,g,...).*

2.6. Análisis de supervivencia

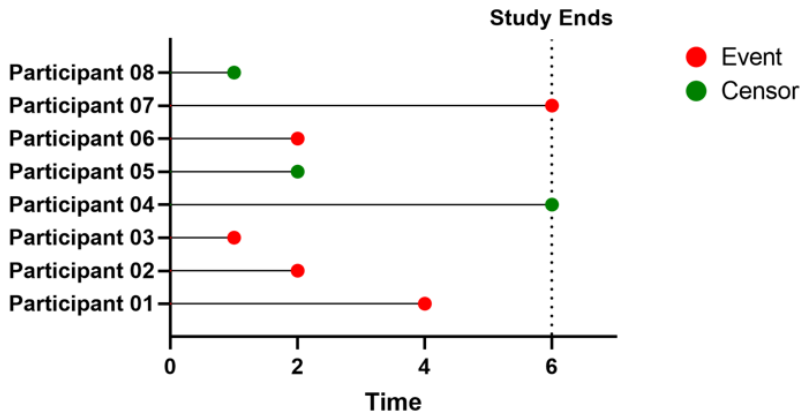
El análisis de supervivencia engloba diferentes técnicas estadísticas para el análisis de datos en los cuales la variable de interés es el tiempo desde el inicio de seguimiento de un individuo hasta que ocurre un evento, ya sea la muerte, recaída, recuperación o cualquier experiencia de interés. Generalmente la variable tiempo es conocida como tiempo de supervivencia, ya que representa el tiempo que un individuo ha sobrevivido durante un proceso de seguimiento; y el evento a su vez como "fallo" ya que suele indicar una experiencia negativa para el individuo. El fin de estos análisis suele ser encontrar factores pronósticos influyentes en la supervivencia, así como evaluar las diferencias de tratamientos. (Kleinbaum & Klein, 2005)

En la mayoría de estos análisis se encuentra el problema de la censura, esta ocurre cuando se tiene información parcial sobre el tiempo de supervivencia, ya sea porque en el tiempo de seguimiento no le ocurre el evento de interés o bien porque se le pierde el seguimiento antes de que ocurra este. Las censuras pueden producirse:

- Por la izquierda, cuando el momento de ocurrencia del evento se desconoce (se produce antes de ingresar al estudio).
- Por la derecha, cuando el evento ocurre después del final del estudio o el individuo abandona el estudio por alguna causa.
- Por intervalos, algunos estudios no se siguen de forma continua la observación de los individuos y el evento ocurre en los periodos entre observaciones, por lo que no se sabe el momento exacto de este y solo se dispone de un intervalo de tiempo.

Al encontrarse censuras en este tipo de datos es por lo que se utiliza este análisis en lugar de modelos de regresión comunes para respuestas continuas como el tiempo.

Ilustración 2. Eventos y censuras en la supervivencia



Nota. El gráfico representa individuos con censuras e individuos en lo que se ha producido el evento de interés. Tomada de www.graphpad.com.

En el análisis de supervivencia, se utiliza T para denotar la variable de respuesta, tiempo hasta que ocurre un evento. El modelo estadístico se define en términos de la función de

supervivencia $S(t)$ que representa la probabilidad de que un individuo sobreviva a un tiempo determinado t. (Harrell , 2015)

$$S(t) = P(T \geq t) = 1 - F(t) \tag{2.25}$$

donde $F(t)$ es la función de distribución acumulativa para T.

El tiempo de supervivencia T puede estar medido como una variable continua o como una discreta por lo que la función de supervivencia cambiaria según el caso:

para T continua:

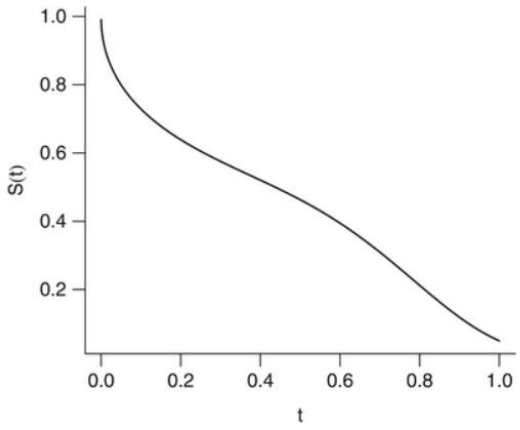
$$S(t) = P(T \geq t) = \int_t^{\infty} f(u)du; \tag{2.26}$$

para T discreta:

$$S(t) = P(T \geq t) = \sum_{t_j \geq t} f(t_j). \tag{2.27}$$

Esta función siempre decrece a medida que t aumenta y para los valores de $t = 0;\ S(t) = 1$ y cuando t tiende a ∞; $S(t) = 0$.

Ilustración 3. Función de supervivencia



Nota. Función de supervivencia. Tomado de *(Harrell , 2015).*

La función de riesgo (conocida como "Hazard Function" o "Hazard Ratio"), $h(t)$ ó $\lambda(t)$, mide el riesgo o probabilidad de que ocurra un evento de interés en un momento específico a un individuo que ha sobrevivido hasta dicho momento, ya que el evento se ha producido antes. Matemáticamente indica la razón entre la probabilidad de que ocurra el evento en un intervalo de tiempo y la longitud de este intervalo entre la probabilidad de que no haya ocurrido el evento antes de este intervalo de tiempo.

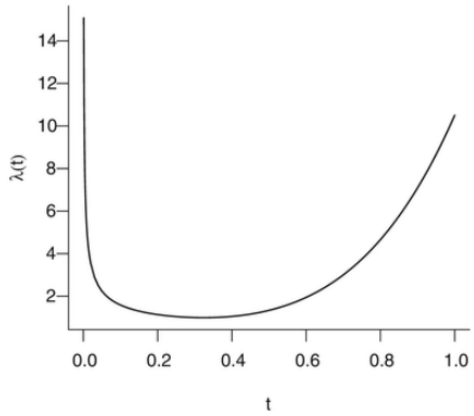
$$h(t) = \lim_{\Delta t \to 0} \frac{P[t < T \leq t + \Delta t | T > t]}{\Delta t} \tag{2.28}$$

También puede ser expresada mediante la función de densidad de T evaluada en $t, f(t) = \frac{\partial S(t)}{\partial t}$, como:

$$h(t) = -\frac{\partial \log S(t)}{\partial t} \qquad (2.29)$$

Al estudiar la función de riesgo se obtienen conocimientos sobre la fuerza del riesgo a lo largo del tiempo ya que cuantifica este, en un momento instantáneo, ya que va variando con el tiempo.

Ilustración 4. Función de riesgo

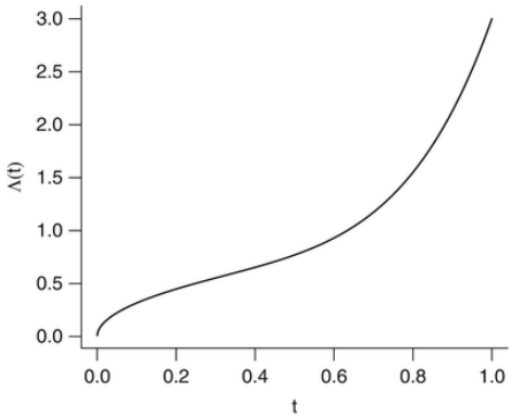


Nota. Función de riesgo. Tomada de (Harrell , 2015).

La función de riesgo acumulado, $H(t)$ ó $\Lambda(t)$, describe el riesgo acumulado hasta el tiempo t y es el negativo del logaritmo de la función de supervivencia.

$$H(t) = -\log S(t) \qquad (2.30)$$

Ilustración 5. Función de riesgo acumulado



Nota. Función de riesgo acumulado. Tomada de (Harrell , 2015).

Esta función creciente a medida que t aumenta, el riesgo acumulado aumenta o se mantiene igual. Su principal propiedad es que el valor esperado de $H(t)$ es la unidad, $E[H(t)] = 1$.

Las principales formas de estimar estas funciones y la supervivencia de un grupo de pacientes se tratan del estimador de Kaplan-Meier y los modelos de riesgos proporcionales de Cox.

2.6.1. Estimador de Kaplan-Meier. Representación de curvas de supervivencia.

El estimador de Kaplan-Meier fue desarrollado por Edward L. Kaplan y Paul Meier en 1958 (Kaplan & Meier, 1958). Es un método no paramétrico para la estimación de la función de supervivencias que se basa en la probabilidad condicionadas y es robusto a los datos con censuras, ya que supone que los individuos censurados se habrían comportado igual que los seguidos hasta el evento. Permite comparar curvas de diferentes grupos a través del test log-rank. (Kleinbaum & Klein, 2005)

Para el cálculo del estimador se realiza la siguiente tabla (*Tabla 2*), ordenando previamente los tiempos de supervivencia de menor a mayor, $t_1 \leq t_2 \leq \cdots \leq t_n$, con sus respectivos números de fallos y censuras producidas. Para cada uno de estos tiempos se calcula la probabilidad de sobrevivir en un tiempo t,

$$S(t) = p_1 \cdot p_2 \cdot \ldots \cdot p_t \tag{2.31}$$

Para una muestra de tamaño n y para cada t_i se definen el número de individuos $n_i$que están en riesgo en el momento anterior a t_i, el numero de eventos en el tiempo d_i y la probabilidad de supervivencia en ese intervalo de tiempo,

$$p_i = 1 - \frac{d_i}{n_i}. \tag{2.32}$$

Tabla 2. Pasos previos al estimador K-M

$\boldsymbol{i}$	$\boldsymbol{t_i}$	$\boldsymbol{n_i}$	$\boldsymbol{d_i}$	$\boldsymbol{p_i}$
1	t_1	n_1	d_1	p_1
2	t_2	n_2	d_2	p_2
…	…	…	…	…
n	t_n	n_n	d_n	p_n

Nota. Tabla para el cálculo del estimador de Kaplan-Meier.

El estimador de Kaplan-Meier se define como:

$$\hat{S}_{KM}(t) = p_1 \cdot p_2 \cdot \ldots \cdot p_t = \left(1 - \frac{d_1}{n_1}\right) \cdot \left(1 - \frac{d_2}{n_2}\right) \ldots \cdot \left(1 - \frac{d_t}{n_t}\right) \tag{2.33}$$

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{2.34}$$

Siendo el intervalo de confianza utilizado en las representaciones graficas para una confianza del 95%:

$$\hat{S}_{KM}(t) \pm z_{0,975} SE\left[\hat{S}_{KM}(t)\right] \tag{2.35}$$

donde

$$SE\left[\hat{S}_{KM}(t)\right] \approx \left[\hat{S}_{KM}(t)\right] \cdot \left[\sum_i \frac{d_i}{n_i(n_i - d_i)}\right]^{\frac{1}{2}} \tag{2.36}$$

Para realizar la comparación de curvas entre diferentes grupos una vez calculadas, se plantea el siguiente contraste de hipótesis, con hipótesis nula referente a que no existen diferencias significativas entre las funciones de supervivencia de los grupos:

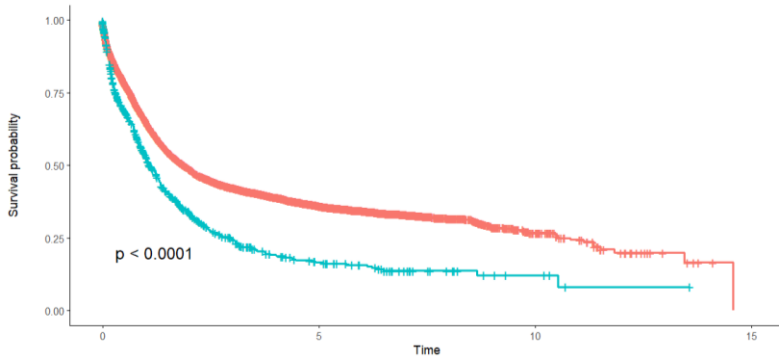
$$\begin{cases} H_0: S_1(t) = S_2(t) \\ H_1: S_1(t) \neq S_2(t) \end{cases} \tag{2.37}$$

Para resolverlo se utiliza el test Log-Rank. Este test fue propuesto por Nathan Mantel (Mantel, 1966) y llamado de este modo por Richard Peto y Julian Peto (Peto & Peto, 1972). Es una prueba no paramétrica que utiliza la chi-cuadrado con *k-1* grados de libertad, siendo k el número de grupos, ya que enfrenta los eventos observados (O_i) frente a los esperados (E_i) si no existiera diferencia entre grupos. Si esta diferencia fuera pequeña no habría razones para afirmar que existiera diferencia, pero en el caso contrario sería lógico afirmar que la diferencia entre grupo no se debe al azar.

Viene dado por:

$$\chi^2_{k-1} \approx \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{2.38}$$

Ilustración 6. Curvas de supervivencia



Nota. Comparación de dos curvas de riesgos con el test Log-Rank. Elaboración propia.

2.6.1.1. Árboles de supervivencia

El estimador de Kaplan-Meier y el test Log-Rank sirven también para la estimación y representación de árboles de supervivencia ("Survival tree"). Estos dividen a la población en grupos más reducidos según las variables más significativas en el test, hasta que el árbol alcanza un numero óptimo de nodos terminales para un p-valor de corte asignado.

Cada ruta desde el primer nodo hasta un nodo terminal especifica una combinación de variables predictoras y sus valores de corte, conducen a un nodo terminal, formando un patrón

de interacción. Cada patrón especifica un subgrupo de individuos con una probabilidad de supervivencia similar dentro del análisis.

2.6.2. Modelo de riesgos proporcionales de Cox

El modelo de riesgos proporcionales de Cox, conocido como regresión de Cox, fue desarrollado por David Cox en 1972 (Cox, 1972). Permite analizar la relación existente, teniendo en cuenta la censura, entre variable explicativas (covariables) y la tasa de riesgo sin hacer ninguna suposición sobre dicha función, por ello es un modelo semiparamétrico ya que hace suposiciones paramétricas sobre el efecto de los predictores en la función de riesgo, pero no hace ninguna sobre la naturaleza de dicha función. (Harrell , 2015)

El modelo se calcula como:

$$h(t, X_1, X_2, \dots, X_p) = h_0(t) \cdot e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p} \quad (2.39)$$

donde la primera componente es el riesgo basal, $h_0(t)$, dependiente del tiempo; y la segunda dependiente de las covariables, β_i.

El modelo recibe el nombre de modelo de riesgos proporcionales ya que, para dos conjuntos diferentes de valores en las covariables, asume que la tasa de riesgo es proporcional entre ellos a lo largo del tiempo. Es decir, asume que los efectos de las covariables son constantes a lo largo del tiempo. Esto permite estimar los ratios de riesgo y cuantificar el efecto de cada covariable en la tasa de riesgo.

Los parámetros del modelo no pueden estimarse con el método de máxima verosimilitud común ya que la función de riesgo es desconocida, por ello Cox argumentó que cuando el modelo de riesgos proporcionales se cumple, se puede obtener una estimación de las covariables β sin estimar $h(t)$, ya que es eliminado de la función de verosimilitud (Cox, 1972). Propuso resolverlo con el método de estimación maximizando la función de verosimilitud parcial, la cual compara las tasas de riesgo de los individuos que experimentan el evento con aquellos que no lo hacen. Esta función se optimiza para obtener los coeficientes de regresión asociados con cada covariable, β_i, que representan la fuerza y dirección de su efecto sobre la tasa de riesgo.

Esta función se representa como:

$$L(\beta) = \prod_{Y_i} \frac{\exp(X_i\beta)}{\sum_{Y_i \geq Y} exp(X_i\beta)} \quad (2.40)$$

La razón de riesgos instantánea (HR), conocida como “Hazard ratio”, es la probabilidad condicional de presentar el evento de interés en el siguiente instante de tiempo y compara los riesgos de dos grupos o individuos diferentes. Se define como:

$$HR = \exp(\beta) \quad (2.41)$$

Si su valor es inferior a la unidad, indica que el efecto de la covariable es protector, mientras que si es superior indica que la covariable es un factor de riesgo.

Despejando de HR podemos obtener el valor del parámetro β:

$$\beta = \log(HR) \tag{2.42}$$

El modelo de COX puede realizarse de forma multivariante y univariante. En el caso del univariante se quiere contrastar si una variable influye o no en el tiempo de supervivencia y si esta influye de manera positiva o negativa en este, mientras que en el modelo multivariante se quiere contras si dicha variable es influyente tanto positiva como negativamente al considerar conjuntamente todas las variables metidas en el modelo.

Para evaluar la bondad de ajuste del modelo uno de los más utilizados son los residuos de Schoenfeld. Fueron desarrollados por David Schoenfeld en 1982 (Schoenfeld, 1982) y se tratan de los valores observados menos los esperados de cada covariante en cada instante de fallo para aquellos individuos no censurados. Su función es proporcionar evidencias de que los efectos de las covariables, β_i, no cambian respecto del tiempo, es decir, son independientes de él, y de haber algún patrón sugiere que el efecto sí varia en el tiempo. Se definen como:

$$r_i^S(t) = X_i(t) - \bar{X}(t)\beta \tag{2.43}$$

donde $X_i(t)$ es el vector con los valores de las covariables para la i-ésima observación en el tiempo t y $\bar{X}(t)$ es el promedio ponderado de los valores de las covariables hasta el tiempo t.

La representación gráfica de estos residuos es de gran ayuda para revelar posibles tendencias o patrones seguidos por las covariables, ya que, si muestran una desviación significativa de la supuesta proporcionalidad, indica que el riesgo no es constante en el tiempo.

3. RESULTADOS

En este apartado se presentarán los diferentes resultados obtenidos tanto en la estadística descriptiva de los datos, así como su interpretación, y los diferentes análisis estadísticos realizados.

3.1. Estadística descriptiva

Para la realización de los análisis descriptivos se utilizará todo el conjunto de pacientes, teniendo en cuenta los datos faltantes al haber recogido y juntado dos bases de datos diferentes.

El estudio se ha realizado en 7796 pacientes de AML, de los cuales el 56% son hombre y 44% mujeres. Los porcentajes de la distribución de género en los pacientes podrían indicar no ser significativo este en la enfermedad, ya que no se observa una diferencia significativa. Posteriormente se comprobará esta suposición mediante un contraste.

Tabla 3. Frecuencia del género

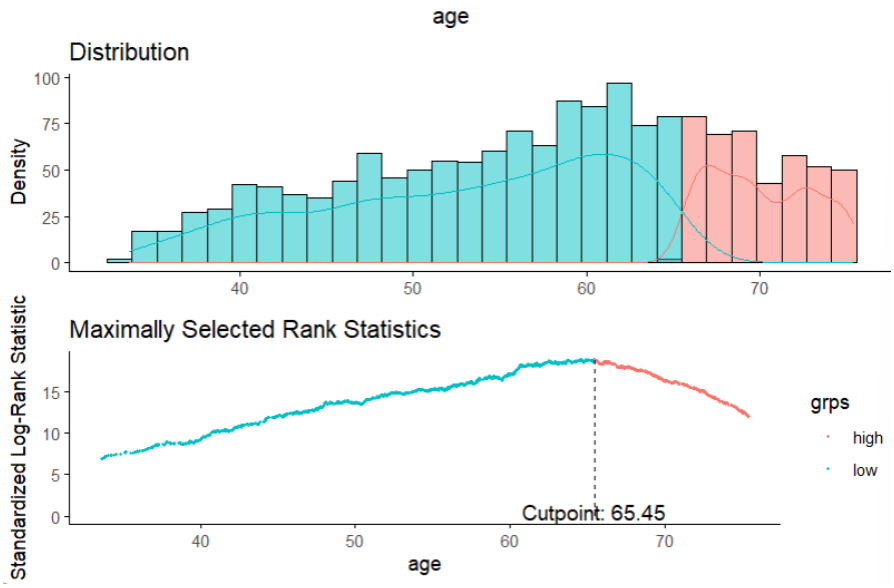
Gender	Male	Female
%	56,03	43,97

Nota. Porcentajes por género.

Como se ha mencionado en la introducción, uno de los modelos más importante actualmente para la estratificación del riesgo en pacientes con AML según diversas alteraciones citogenéticas y moleculares es el European LeukemiaNet (ELN) y por ello es una de las principales variables a estudio en nuestros datos. Se han obtenido medianas y frecuencias de diferentes variables a estudio según las 3 categorías de riesgo de ELN.

Previamente la variable que recoge la edad de los pacientes, distribuida en un intervalo entre los 16 hasta los 100 años, ha sido recodificada en dos grupos de edad según un punto de corte óptimo, 65,4 años, según cómo están distribuidos los valores. Siendo las categorías de esta nueva variable *"high*" para valores superiores a este punto y *"low"* para los inferiores.

Gráfica 1. Recodificación en grupos de la edad



Nota. Punto de corte óptimo para recodificar la variable edad en dos grupos.

Se puede observar en la Tabla 4 como la edad, en media, es superior en el peor riesgo de la enfermedad, 65 años; así como en los grupos de edad el grupo más joven tiene mayor

frecuencia en los riesgos favorable e intermedio 76-77% respecto a 55% del riesgo adverso; mientras la categoría *"high"* aparece con mayor frecuencia en el peor de los riesgos 55% frente a 22-23%. Esto afirma lo encontrado en la teoría de la enfermedad, la edad puede considerarse un factor de riesgo ya que los pacientes más jóvenes tienden a tener mejor pronóstico, mientras que a mayor edad peor pronóstico y mayor riesgo.

Tabla 4. Valores observados para los grupos de ELN

Variable/Eln_2017		Favourable	Intermediate	Adverse	Total
Age	**<65,4**	657 (76,04)	358 (77,66)	438 (55,84)	1453 (68,76)
	>65,4	207 (23,96)	103 (22,34)	350 (44,16)	660 (31,24)
	median	56,1(16-86,4)	56,6 (15,9-100)	65,5(16,3-91,1)	59,21 (15,9-100)
Gender	**M**	419 (48,50)	246 (53,36)	519 (65,86)	1184 (56,03)
	F	445 (51,50)	215 (43,64)	269 (34,14)	926 (43,97)
Median	**wbc**	28,3 (0,3-456)	9,9 (0-386)	9,95 (0,4-430)	16,2 (0-456)
	hb	9,2 (0,72-139)	9,7 (3,5-144)	9,2 (0,8-176)	9,3 (0,72-176)
	plt	54 (2-1523)	63 (2-2013)	58,5 (2-937)	57,0 (2-2013)
%	**bm_blasts**	70 (0-100)	58,7 (0-100)	53 (0-100)	60,0 (0-100)
Type	**pAML**	810 (93,75)	383 (83,08)	555 (70,43)	1748 (82,73)
	sAML	54 (6,25)	78 (16,92)	233 (29,57)	365 (17,27)
Total		864	461	788	2113

Nota. Medianas y porcentajes de las variables según las categorías de la variable eln_2017 clasificando el riesgo.

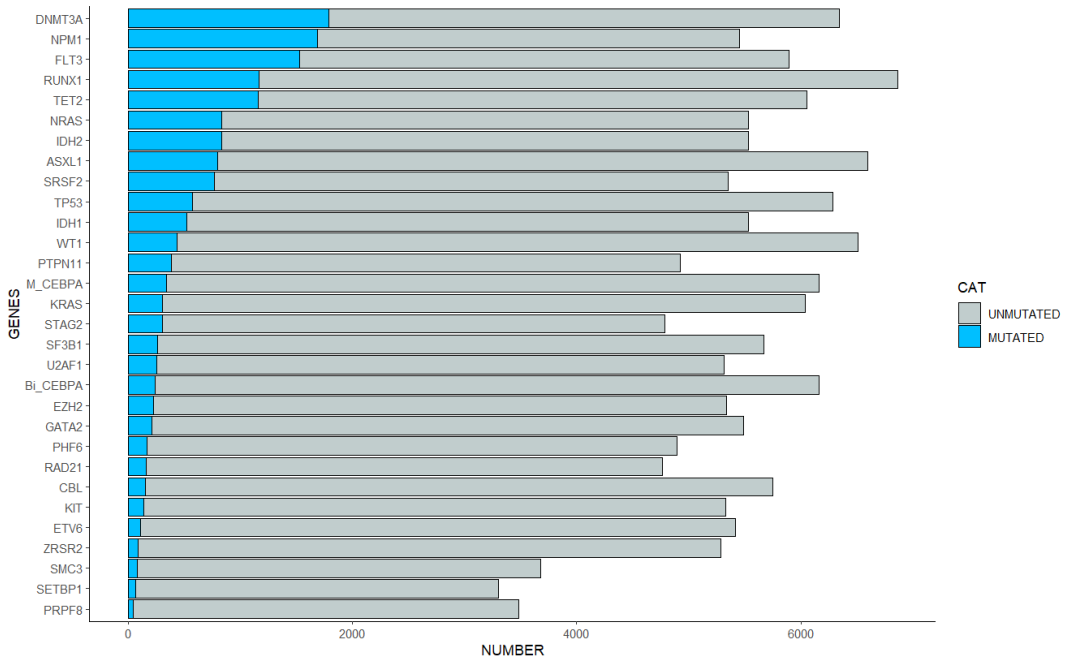
En cuanto a los biomarcadores estudiados, glóbulos blancos, hemoglobina, plaquetas y blastos, no se han encontrado grandes diferencias entre los grupos de riesgo ya que hay un amplio intervalo de valores en todos ellos. Aun así, se puede destacar que tanto en media los glóbulos blancos como el porcentaje de blastos en sangre son superiores en el grupo de riesgo favorable.

El tipo de ocurrencia de AML se clasifica como primaria (pAML) si surgía como nueva sin antecedentes, o secundaria (sAML) si venía derivada de otra o de un tratamiento previo. Estudios previos han indicado que la primaria tiene un mejor pronóstico y menor riesgo que la secundaria como se ha podido ver en la introducción. En la Tabla 4 se puede observar que en nuestros datos se ha confirmado esto, ya que en la pAML el mayor porcentaje de ocurrencia se encuentra en el grupo de riesgo favorable, 94% frente a un 70% en el riesgo desfavorable; mientras que la sAML presenta un mayor porcentaje en el peor grupo de riesgo,30% frente a un 6% en el mejor.

Las alteraciones citogenéticas y las mutaciones de genes tienen una gran importancia en la aparición y desarrollo de la AML. Por ello, se ha estudiado primero el número de mutaciones producidas en los genes en todos los pacientes.

Como se puede observar en el diagrama de barras Gráfica 2 el gen que mayor número de mutaciones tiene entre los pacientes es el DNMT3A, seguido de NPM1, FLT3 y RUNX1. Coincidiendo estos con los genes más relevante en el transcurso de la enfermdad y siendo claves para la estratificación del riesgo.

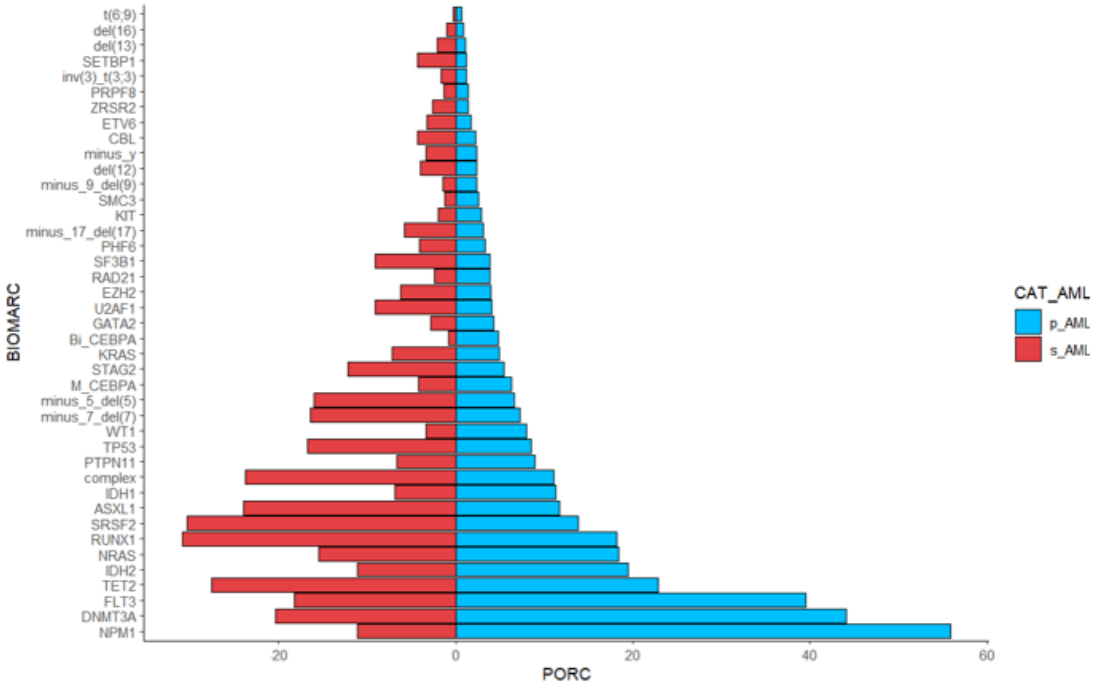
Gráfica 2. Gráfico de barras de las mutaciones de genes



Nota. Número de pacientes según la mutación de cada gen.

Posteriormente, se ha estudiado la frecuencia de aparición de las alteraciones citogenéticas como de las mutaciones de genes para las dos categorías de AML según su ocurrencia (Gráfica 3). Para las AML primarias el mayor porcentaje de ocurrencia son las mutaciones de los genes NPM1, DNMT3A y FLT3, coincidiendo estos con los de mejor pronóstico o pronóstico intermedio según la ELN. En cambio, para las AML secundarias aparecen con mayor frecuencia las mutaciones de RUNX1, SRSF2, TET2, ASLX1 y la posesión de cariotipo complejo, coincidiendo con las anomalías genéticas asignadas a un riesgo desfavorable.

Gráfica 3. Gráfico de barras por categorías de AML con alteraciones genéticas



Nota. Porcentaje de mutación por gen en los diferentes tipos de aparición de AML.

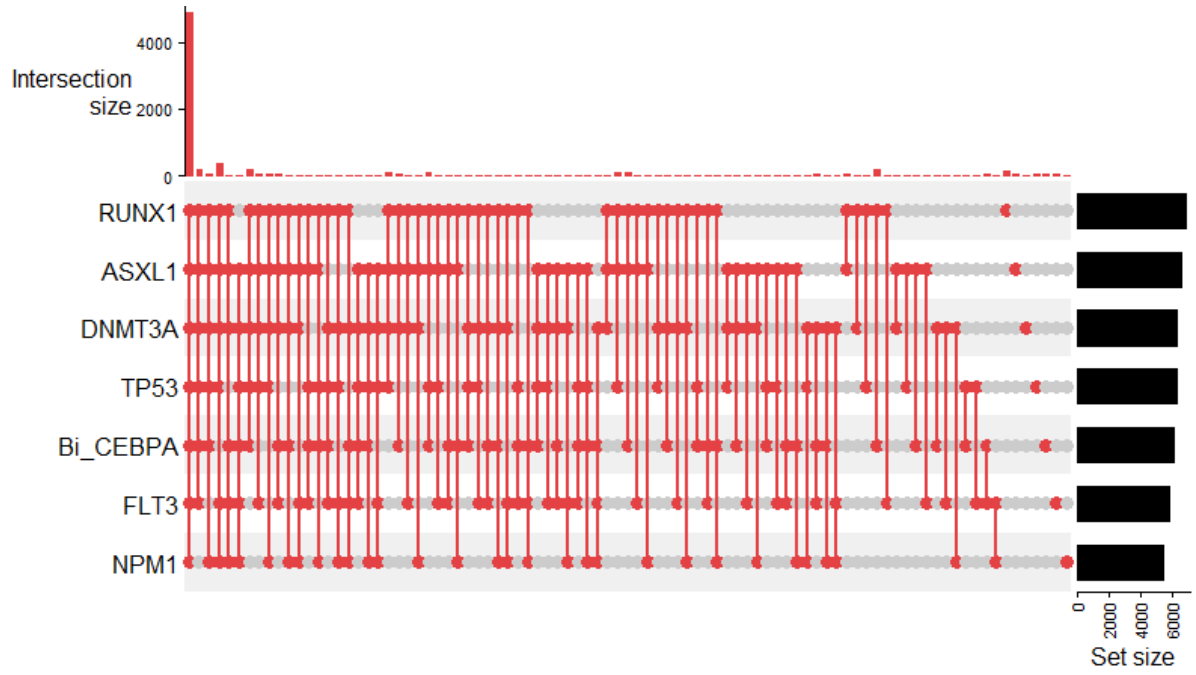
Por tanto, se confirmaría que el tipo de AML secundario tiene un peor pronóstico y mayor riesgo que la AML primaria.

Para la obtención de los siguientes resultados se han selccionado los genes más relevantes en la enfermedad según los resultados obtenidos anteriormente, la teoría y la importancia en la estratificación del riesgo. Estos son: ASLX1, Bi-CEBPA, DNMT3A, FLT3, NPM1, RUNX1 y FLT3.

Para poder evaluar el riesgo de las mutaciones de genes, es interesante saber cuáles tienden a aparecer de forma simultanea en los pacientes.

En la Gráfica 4 se puede observar cómo la intersección de genes que más veces ocurre contiene a los siete seleccionados anteriormente, seguida de la combinacion formada por todos ellos menos el gen Bi-CEBPA. Coincidiendo éste con uno de los de menor importancia a la hora de estratificar el riesgo y de mejor pronóstico.

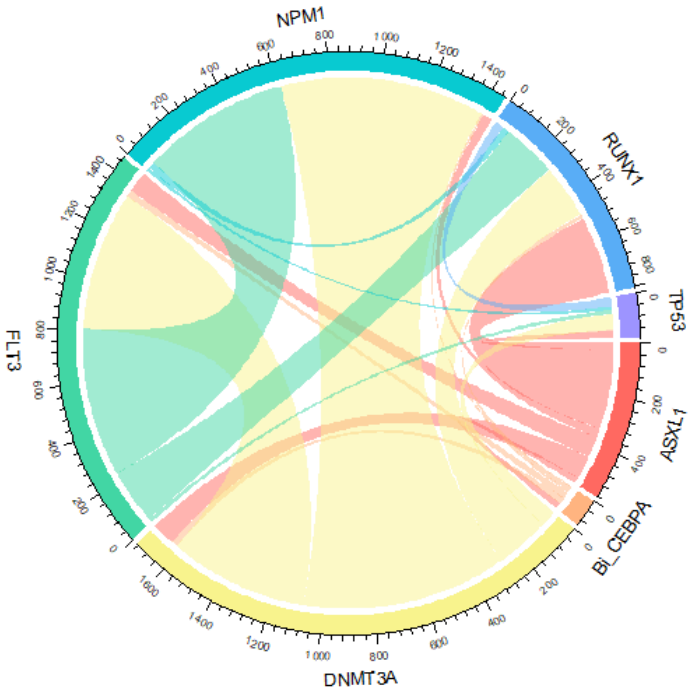
Gráfica 4. Upset plot con genes



Nota. Frecuencia de apariciones conjuntas e individuales de los genes.

Siguiendo con el estudio de las intersecciones entre genes, para observar con cuales de ellos tienen una coocurrencia más alta entre sí, se ha elaborado el siguiente gráfico. (Gráfica 5)

Gráfica 5. Chord diagram de genes



Nota. Coocurrencias de los genes importantes en el desarrollo de la AML.

Entre los genes del grupo de riesgo desfavorable (RUNX1, ASLX1, TP53) se observa que el gen ASLX1 muta mayormente con los genes de este grupo de riesgo y RUNX1 con FLT3. Entre los de riesgo intermedio (NPM1 con FLT3) se puede ver que el gen NPM1 muta conjuntamente en su mayoría con el gen DNMT3A y FLT3. Confirmando dicha intersección en el riesgo intermedio. Por último, en los de riesgo favorable (BI-CEBPA y NPM1) no se observa ninguna mutación destacable en el gen BI-CEBPA.

Cabe destacar que se observa que si un gen de riesgo favorable muta conjuntamente con uno de riesgo desfavorable, en la estratificación del riesgo toma mayor importancia este último.

Tras estudiar los posibles alteraciones citogenéticas, biomarcadores, factores clínicos y factores de riesgo, se llevarán a cabo diferentes contrastes para ver la significancia de estos y poder realizar diferentes modelos de supervivencia.

3.2. Contrastes de hipótesis

Antes de la realización de los contrastes de hipótesis para comprobar si existen diferencias significativas entre las categorías de las variables tipo de AML y ELN con diferentes variables, se ha contrastado la normalidad de los datos en las variables numéricas y la homogeneidad de varianzas en los diferentes grupos de las variables a estudio.

En el contraste de normalidad realizado con el test de Shapiro-Wilk, se ha obtenido que ni la edad, ni los glóbulos blancos (wbc), plaquetas (plt), blastos (bm_blasts) ni hemoglobina (hb) siguen una distribución normal, rechazando la hipótesis nula de normalidad con un p-valor significativo en todos los casos de $2{,}2^{-16}$.

Para contrastar la homogeneidad de las varianzas, como las variables no siguen una distribución normal, se ha utilizado el test de Levene ya que es menos sensible que el test de

Bartlett ante la falta de normalidad. Obteniendo los siguientes p-valores en los contrastes de las variables numéricas con AML_Type y eln_2017 (Tabla 5):

Tabla 5. P-valores en test de Levene

	age	wbc	plt	bm_blasts	hb
AML_Type	$4{,}705e^{-12}$	$1{,}528e^{-5}$	0,121	0,942	0,014
eln_2017	0,014	$6{,}678e^{-7}$	0,016	$1{,}241e^{-5}$	0,027

Nota. P-valores del contraste de homocedasticidad con el test de Levene para las variables AML_Type y eln_2017.

Para la variable del tipo de AML se obtienen p-valores significativos para las variables numéricas edad, glóbulos blancos y hemoglobina, por lo que se rechaza la homogeneidad de varianzas de estas entre las categorías de esta variable. En cambio, para la variable de estratificación del riesgo ELN se obtienen todos los p-valores significativos por lo que se rechaza la hipótesis nula de homogeneidad de varianzas en las categorías de esta para todas estas variables numéricas.

Una vez contrastadas la normalidad y la homocedasticidad, se estudian las posibles diferencias entre algunas variables y las categorías del tipo de AML según el comienzo de la enfermedad, teniendo en cuenta los resultados obtenidos en la estadística descriptiva.

En primer lugar, se han contrastado las variables clínicas numéricas, para comprobar si existen diferencias entre las distribuciones de las dos categorías de AML, mediante el test no paramétrico de Wilcoxon-Mann-Whitney, obteniendo los siguientes p-valores. (Tabla 6)

Tabla 6. P-valores test Wilcoxon-Mann-Whitney

	age	wbc	plt	bm_blasts	hb
AML_Type	$2{,}200e^{-16}$	$1{,}699e^{-5}$	0,449	$2{,}200e^{-16}$	$2{,}200e^{-16}$

Nota. P-valores del contraste no paramétrico Wilcoxon-Mann-Whitney en la variable AML_Type.

A excepción de las plaquetas, se obtienen p-valores altamente significativos tanto para la edad, glóbulos blancos, blastos y hemoglobina, por lo que se pueden considerar sus distribuciones diferentes en las categorías sAML y pAML, y por tanto, ser relevantes en el estudio de riesgo y clasificación de los pacientes.

Para el caso de la variable clínica que indica el género de los pacientes y los biomarcadores genéticos estudiados en la estadística descriptiva (ASLX1, Bi-CEBPA, DNMT3A, FLT3, NPM1, RUNX1 y TP53) se ha llevado a cabo el contrate mediante el test Chi-cuadrado. Para todos ellos se ha obtenido un p-valor altamente significativo de $2{,}200e^{-16}$, por lo que se rechaza la hipótesis nula de independencia y se considera que estas variables son relevantes en la categorización de los pacientes de sAML y pAML.

Tras ello se llevan a cabo los contrastes de anova no paramétrico mediante el test de Kruskal-Wallis de estas variables con la ELN de estratificación del riesgo (eln_2017) en favorable, intermedio y desfavorable para ver si existen diferencias entre las distribuciones de cada variable numérica en las 3 categorías. *(*Tabla 7)

Tabla 7. P-valores en test de Kruskal-Wallis

	age	wbc	plt	bm_blasts	hb
eln_2017	$2{,}200e^{-16}$	$2{,}200e^{-16}$	0,077	$2{,}200e^{-16}$	$3{,}706e^{-5}$

Nota. P-valores del contraste no paramétrico Kruskal-Wallis en la variable eln_2017.

Se han encontrado diferencias significativas entre las distribuciones de los tres grupos de riesgo para la edad, glóbulos blancos, blastos y hemoglobina al obtener p-valores altamente significativos.

Para estudiar entre qué grupos de los tres se encuentran estas diferencias se han realizado los contrastes post-hoc con comparaciones dos a dos con la prueba de Wilcoxon-Mann-Whitney. (Tabla 8)

Tabla 8. Post-hoc dos a dos Wilcoxon-Mann-Whitney

	age	wbc	bm_blasts	hb
fav-int		*	*	*
int-adv	*		*	*
adv-fav	*	*	*	

Nota. Los * indican las diferencias significativas con un p-valor ≤ 0,05 entre las categorías de ELN.

Para estudiar si influyen en la clasificación de las tres categorías de riesgo el género de los pacientes y las mutaciones de los genes seleccionados previamente (ASLX1, Bi-CEBPA, DNMT3A, FLT3, NPM1, RUNX1 y TP53), se ha realizado un contraste con el test Chi-cuadrado. Se ha obtenido unos p-valores altamente significativos de entre $2{,}200e^{-16}$ y $1{,}174e^{-7}$ en todos los contrastes, por lo que nuevamente se rechaza el supuesto de independencia de estos con la variable de la ELN y se puede afirmar que son factores influyentes para la estratificación del riesgo en pacientes con AML.

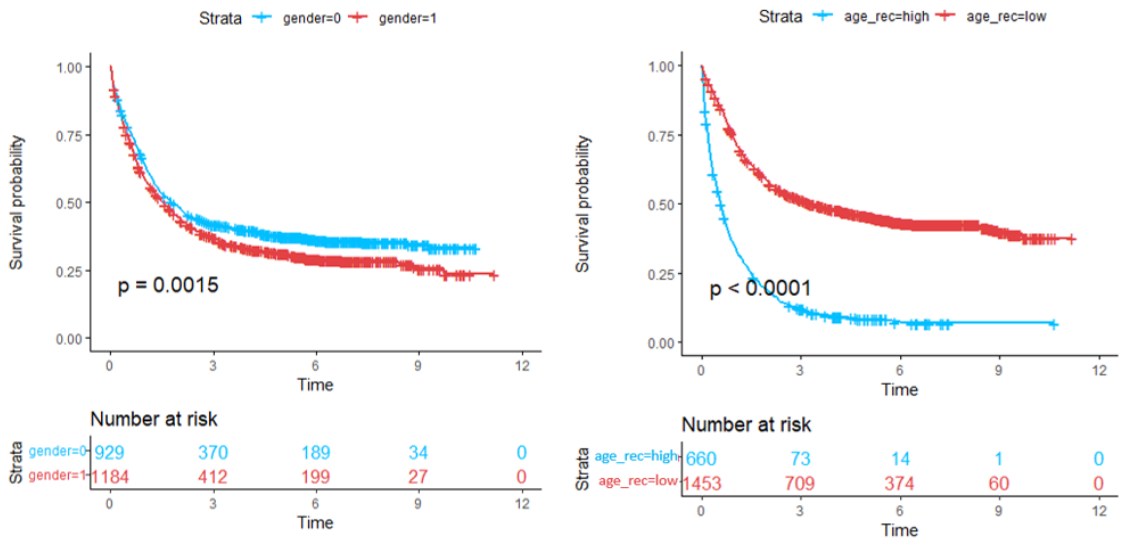
3.3. Análisis de supervivencia

3.3.1. Curvas de supervivencia

Con las variables clínicas y los biomarcadores previamente contrastados, se han llevado a cabo análisis de supervivencia no paramétricos con el estimador de Kaplan-Meier, obteniendo diferentes curvas y comparaciones de éstas con el test Log-Rank.

Como se pudo observar en los análisis descriptivos tanto la edad como el género de los pacientes se podían considerar factores de riesgo ya que había diferencias significativas tanto en la estratificación del riesgo como en la clasificación de la AML. Por tanto, hay razones para pensar que puedan existir diferencias significativas entre las categorías del género y los grupos de edad en la supervivencia de los pacientes.

Gráfica 6. Comparación de curvas de supervivencia de género y edad

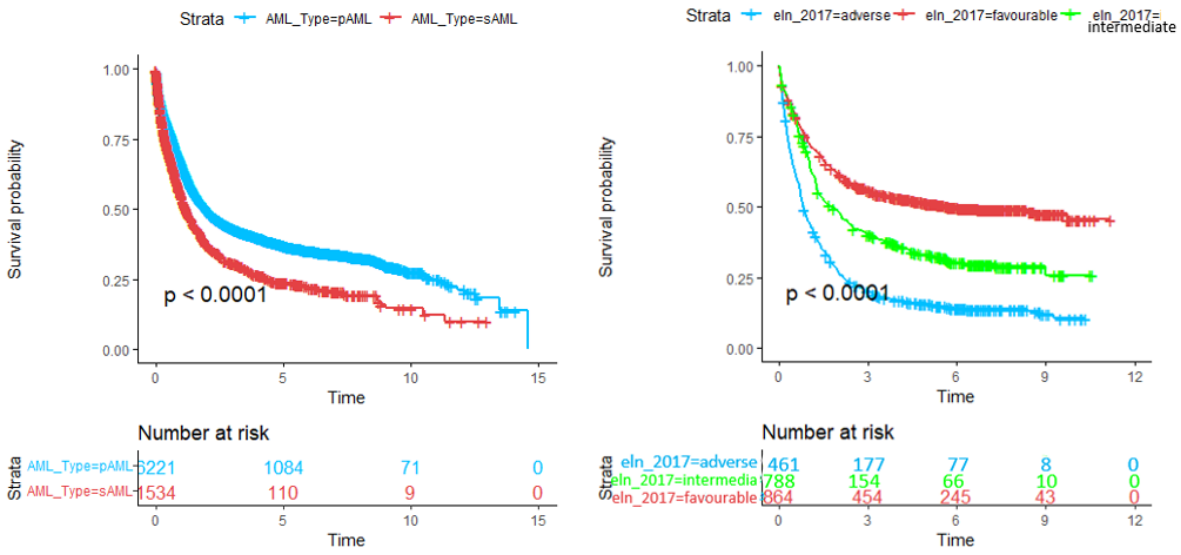


Nota. Curvas K-M de género y edad.

Como se puede ver en la Gráfica 6, en ambas variables existen diferencias significativas entre las dos curvas de cada categoría ya que se obtienen con el test de Log-Rank p-valores altamente significativos en los dos casos. En la variable género se aprecia que la diferencia es favorable en las mujeres ya que tienen una mejor supervivencia que los hombros, por tanto, los estos tienden a tener un peor pronóstico y mayor riesgo en la enfermedad. En el caso de la edad existe una gran diferencia en la supervivencia entre pertenecer al grupo de edad más joven (<65,4 años), mejor supervivencia, y al mayor (>65,4 años). Por lo que se confirmaría lo afirmado en la teoría de que, a mayor edad, la probabilidad de recuperación disminuye y aumenta la de recaída y muerte.

Para corroborar que las AML secundarias tienen peor pronóstico que las primarias y afirmar la diferencia de supervivencia según la estratificación de riesgo de la ELN, se elaboran curvas con estas variables.

Gráfica 7. Comparación de curvas de supervivencia de tipo de AML y ELN

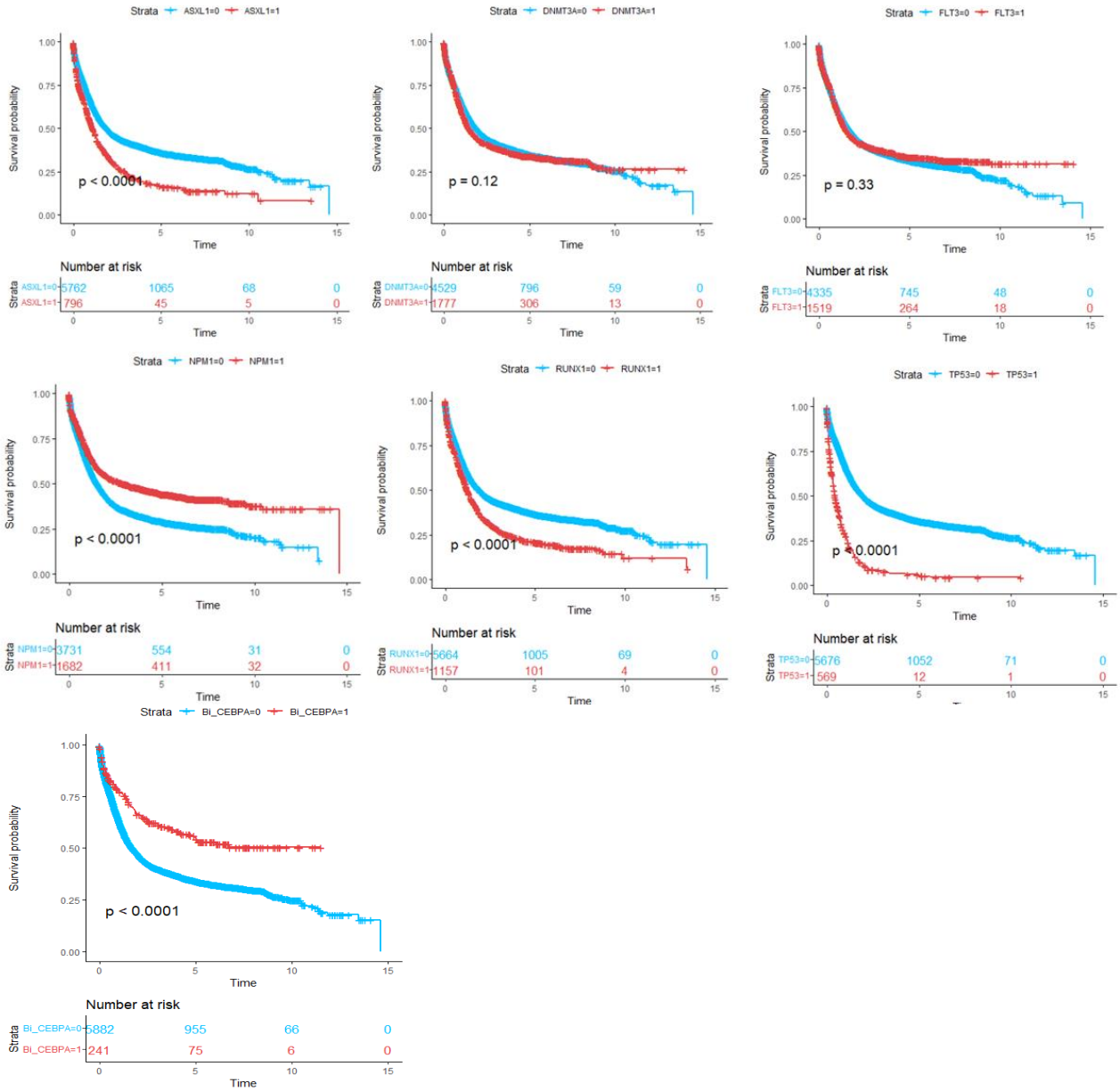


Nota. Curvas K-M de tipo de AML y clasificación ELN 2017.

Como era de esperar, se obtienen p-valores altamente significativos en el test (Gráfica 7) y se puede confirmar que existen diferencias significativas en la supervivencia entre los diferentes grupos. Los pacientes con AML secundaria tienen una peor supervivencia que lo que padecen una primaria, y efectivamente, en la ELN las AML clasificadas como de riesgo adverso por sus alteraciones citogenéticas son la de peor supervivencia y las de mejor las categorizadas como favorables.

Entre los genes seleccionados previamente (ASLX1, BI-CEBPA, DNMT3A, FLT3, NPM1, RUNX1 y TP53) interesa estudiar la supervivencia de los pacientes para poder sacar conclusiones acerca de si sus mutaciones afectan de forma favorable o adversa en la enfermedad.

Gráfica 8. Comparación de curvas de supervivencia de genes



Nota. Curvas K-M para los genes ASLX1, Bi-CEBPA,DNMT3A, FLT3, NPM1, RUNX1 y TP53.

En el test se obtienen p-valores significativos en las curvas de los genes ASLX1, BI-CEBPA, NPM1, RUNX1 y TP53 por lo que la supervivencia en los pacientes si experimenta una diferencia en ellos, mientras que las curvas de DNMT3A y FLT3 de los pacientes que tiene sus mutaciones y de los que no, no experimentan diferencias.

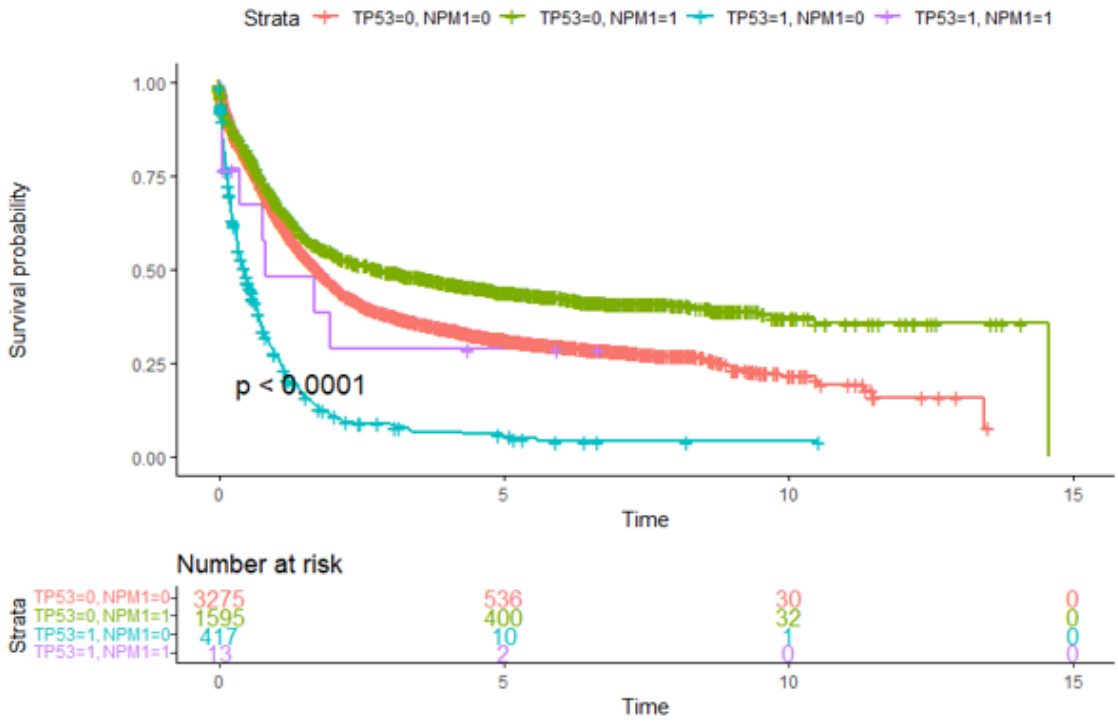
Para los genes BI-CEBPA y NPM1 se obtienen diferencias favorables y altamente significativas en la supervivencia de los pacientes que padecen estas mutaciones, por lo que se les puede considerar un factor favorable en el curso de la enfermedad.

En el caso de los genes ASXL1, RUNX1 y TP53 ocurre, al contrario, las diferencias altamente significativas encontradas entre las curvas son desfavorables en los pacientes que padecen estas mutaciones, por lo que estos genes pueden considerarse factores de riesgo.

Tanto la clasificación de BI-CEBPA y NPM1 como favorables en la AML, como de ASXL1, RUNX1 y TP53 como desfavorables y factores de riesgo, coincide con la clasificación de las anomalías genéticas en los grupos de riesgo de la ELN 2017.

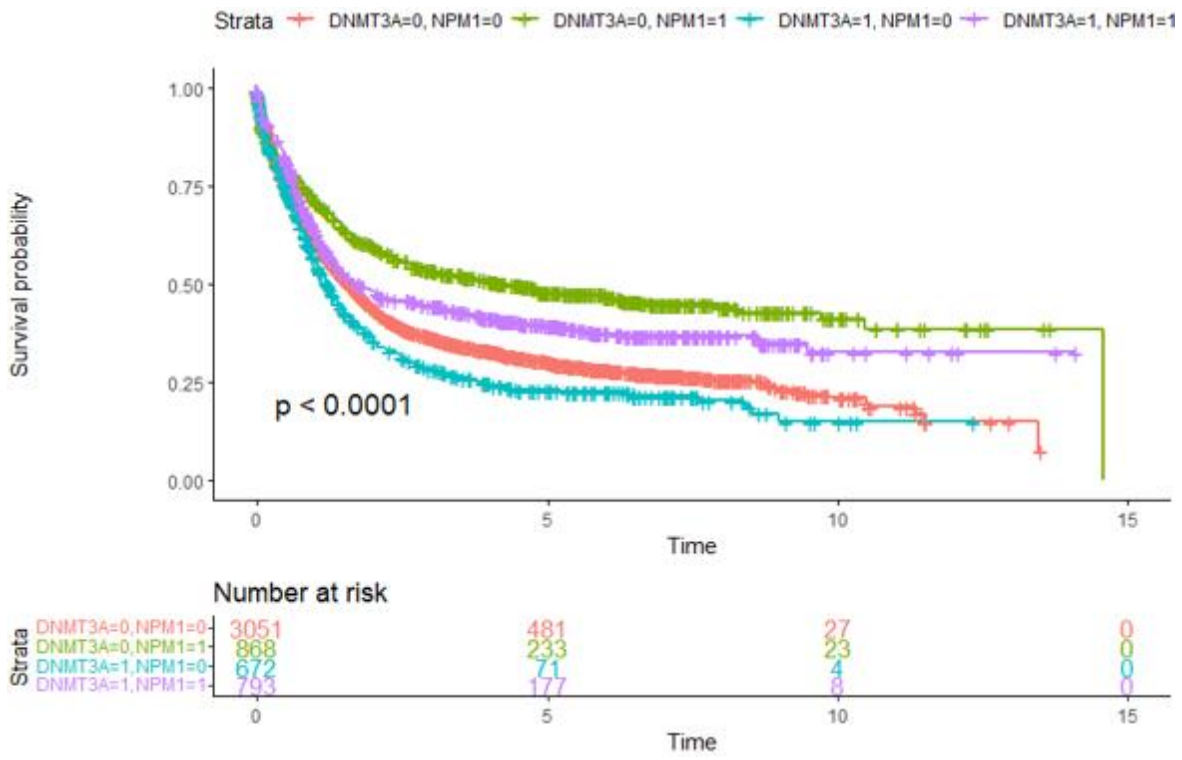
Se ha querido comprobar qué ocurre y qué riesgo prima en la supervivencia de los pacientes cuando un gen favorable muta simultáneamente con uno desfavorable, como NPM1 conTP53; así como cuándo uno favorable muta con uno neutro, como NPM1 con DNMT3A. (Gráfica 9)

Gráfica 9. Comparación de curvas de supervivencia dos genes



Nota. Curvas K-M para las mutaciones conjuntas TP53-NMP1.

Gráfica 10. Comparación de curvas de supervivencia dos genes



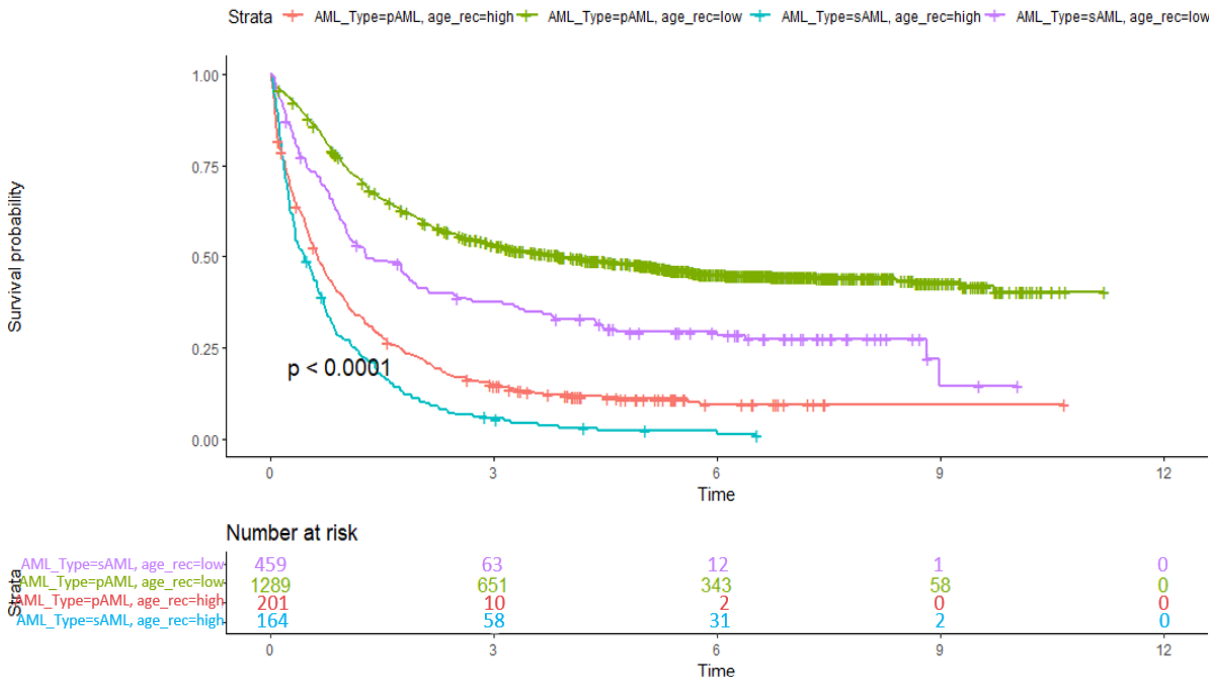
Nota. Curvas K-M para las mutaciones conjuntas DNMT3A-NPM1.

En las curvas de TP53 con NPM1 se encuentran diferencias altamente significativas entre ellas, obteniendo la peor supervivencia cuando TP53 muta sin NPM1 y la mejor en el caso contrario, y se encuentra peor supervivencia cuando mutan conjuntamente respecto de cuando ninguna muta. Por lo que puede decirse que en presencia de un gen favorable y uno de riesgo prima en la estratificación del riesgo el desfavorable, en este caso TP53.

En el caso de un gen no significativo en la AML, como DNMT3A, y uno favorable en el curso de la enfermedad, como NPM1, se obtienen diferencias en las curvas altamente significativas también. Teniendo la mejor supervivencia cuando el gen favorable aparece en ausencia del neutro y la peor cuando DNMT3A muta sin NPM1. Por tanto, podría decirse que se ha observado que si NPM1, a pesar de ser favorable, va acompañado de otro ya sea neutro o no clasificado, deja de ser tan bueno en el transcurso de la AML.

Por último, se ha querido contrastar la supervivencia del tipo de AML junto con los grupos de edad para comprobar si realmente las AML secundarias son peor que las primarias independientemente de la edad o si también la edad afecta positiva o negativamente a estas. (Gráfica 10)

Gráfica 11. Comparación de curvas de supervivencia con edad y tipo de AML



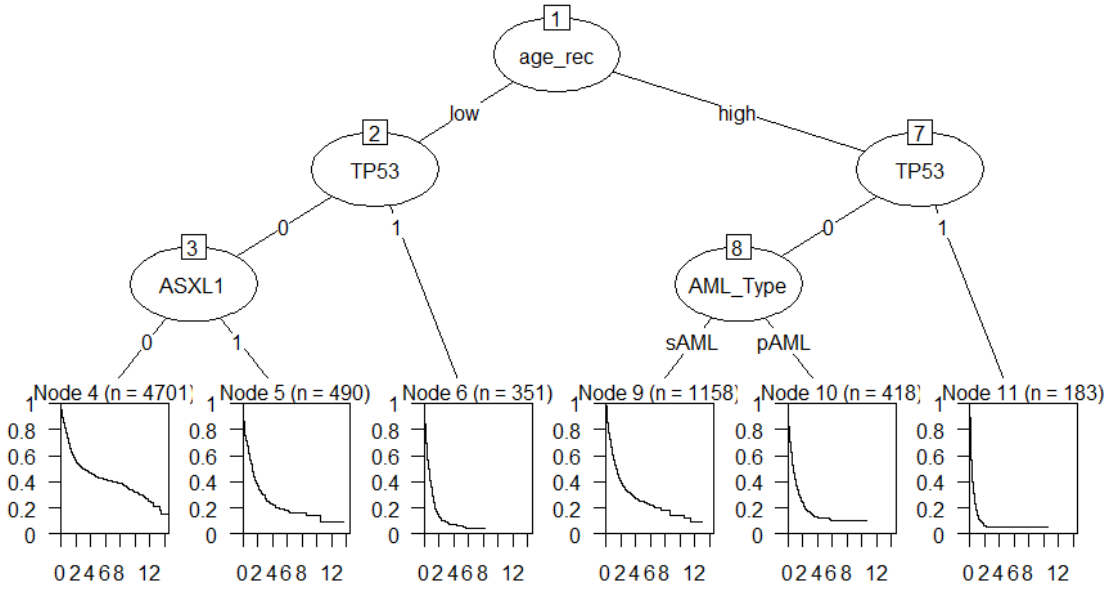
Nota. Curvas K-M para tipo de AML junto con los grupos de edad.

Se obtiene en el test Log-Rank un p-valor altamente significativo, por lo que hay diferencias entre las curvas de los diferentes grupos. Obteniéndose la peor supervivencia para las AMl secundarias acompañadas en los pacientes de mayor edad seguida de las AML primarias en el mismo grupo de edad, y la mejor supervivencia en el grupo más joven con AML primarias. Por lo que podría afirmarse que ante padecer una AML primaria o secundaria o ser un paciente más mayor, se considera mayor factor de riesgo la edad del paciente, ya que como se ha comprobado, independientemente del tipo de AML, si el paciente pertenece al grupo de edad mayor tienen peor supervivencia.

Mediante Kaplan-Meier se ha estimado también las curvas de supervivencia dentro del árbol de supervivencia (“Survival tree”), el cual divide a los pacientes en grupos más reducidos según las variables más significativas del test Log-Rank (Gráfica 11).

Como se ha dicho el párrafo previo, la edad es el factor de riesgo que más prima frente a los tipos y riesgos del AML, por lo que se quiere comprobar si realmente en el árbol prevalece frente a las demás variables consideradas significativas en la supervivencia y cuáles son las más importante para la clasificación de los pacientes en grupos de riesgo.

Gráfica 12. Árbol de supervivencia



Nota. Árbol de supervivencia con 11 nodos y 6 grupos similares.

El algoritmo del árbol de supervivencia ha identificado para un p-valor de corte significativo de al menos 0,003 a 6 grupos diferentes de pacientes. Siendo el nodo principal la edad de los pacientes la cual los divide en dos ramas, en una a los de menor edad y en otra a los de mayor. Siguiendo por la rama de los de menor edad, lo más significativo es el gen TP53 teniendo dentro de esta rama la peor supervivencia los que presenta su mutación (nodo 6), seguida de los que no la presentan, pero sí tiene el gen ASXL1 mutado (nodo 5). Por la rama de los de mayor edad, también lo siguiente más significativo es el gen TP53 y los que peor supervivencia presentan de todo el árbol son los que tienen su mutación y pertenecen a esta rama (nodo 11).

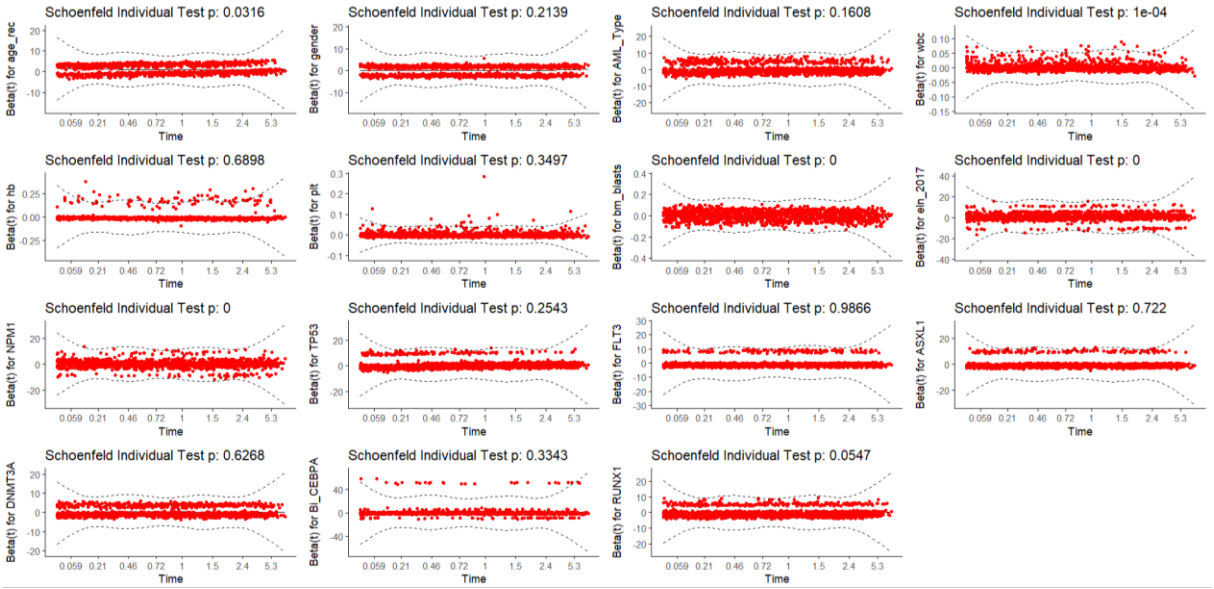
Por lo tanto, el peor factor de riesgo clínico en AML es la edad, para los pacientes de más de 65 años, seguido independientemente de la edad del paciente presentar la mutación del gen TP53.

3.3.2. Modelo de riesgos proporcionales de Cox

Se realizarán los modelos de Cox univariante para ver el efecto tienen cada una de las variables clínicas y biomarcadores en la supervivencia de los pacientes, y partir de los resultados obtenidos se realizará el modelo univariante.

Previamente se llevan a cabo los residuos de Schoenfeld en un modelo que incluye todas las variables a estudiar: características del paciente (grupo de edad y género), variables clínicas (recuento de glóbulos blancos, plaquetas, blastos y hemoglobinas), tipo de AML, clasificación en la ELN y biomarcadores (genes significativos en Kaplan Meier ASXL1, BI-CEBPA, NPM1, TP53, RUNX1).

Gráfica 13. Residuos de Schoenfeld

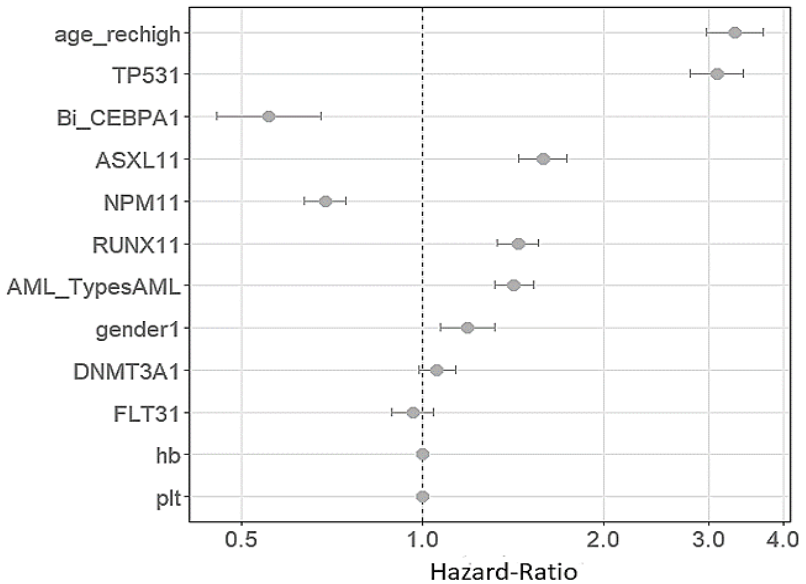


Nota. Residuos de Schoenfeld para la introducción de variables al modelo.

Se obtienen (Gráfica 12) p-valores no significativos para las variables género del paciente, tipo de AML, recuento de hemoglobina, plaquetas y genes tP53, FLT3, ASXL1, DNMT3A, BI-CEBPA y RUNX1. Por lo que estas variables interesa estudiarlas en el modelo de Cox, aunque además se añadirán la edad, ya que como hemos visto en los anteriores análisis es un importante factor de riesgo.

Se realiza el análisis univariante con estas variables y se obtienen p-valores altamente significativos para el grupo de mayor edad, para el género hombre, el tipo de AML secundario, y los genes mutados TP53, BI-CEBPA, ASLX1, NPM1 y RUNX1. Se obtiene la siguiente representación del modelo con los Hazard Ratio (Gráfica 13):

Gráfica 14. Riesgos del modelo univariante



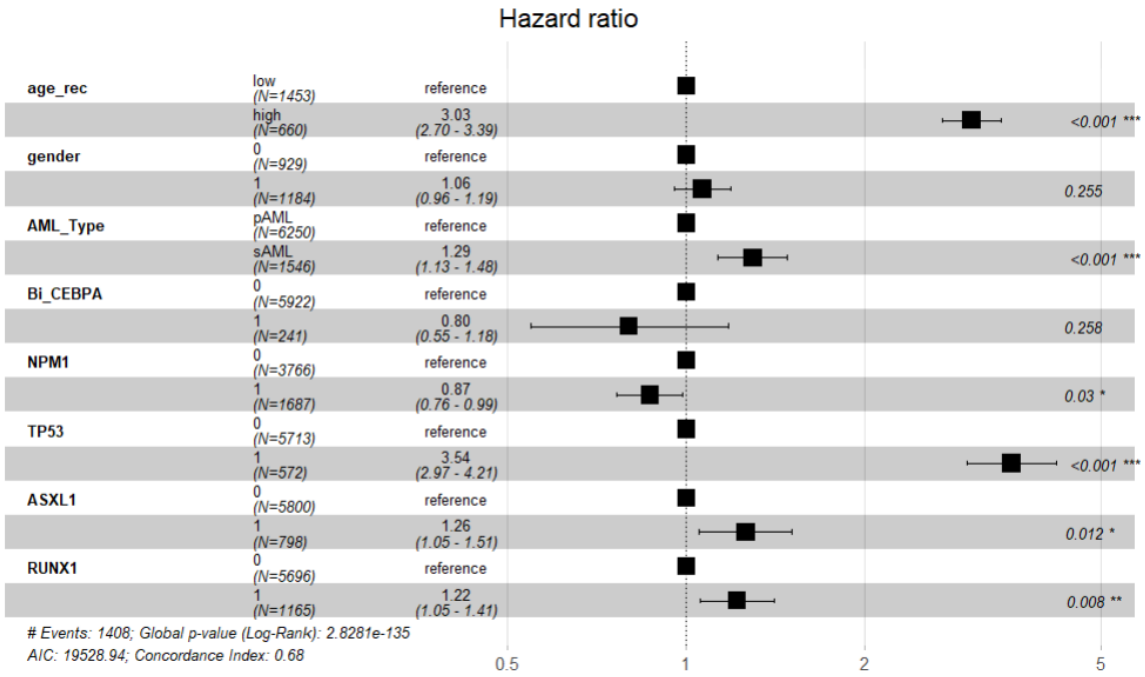
Nota. Hazard- Ratios modelo de riesgos proporcionales de Cox univariante.

Como se puede observar en el gráfico, tener más de 65 años supone entre 3 y 4 veces más de riesgo respecto de un paciente más joven, al igual que presentar el gen TP53 respecto a no tener la mutación. Padecer una AML secundaria presenta casi el doble de riesgo frente a las primarias, igual que presentar los genes ASXL1 y RUNX1. En cambio, presentar los genes BI-CEBPA y NPM1 reduce el riesgo de la AML en casi la mitad frente no tener estas mutaciones.

Con las variables significativas en este modelo, se ha llevado a cabo el multivariante, el cual contrasta la influencia de cada una de ellas al considerarse conjuntamente con el resto.

Se obtienen p-valores significativos para las variables edad, tipo de AML y los genes NPM1, TP53, ASLX1 y RUNX1 con los siguientes Hazard Ratio. (Gráfica 14)

Gráfica 15. Riesgos del modelo multivariante



Nota. Hazard Ratios del modelo de riesgos proporcionales de Cox multivariante.

Al considerarse conjuntamente con el resto de las variables, ser hombre deja de ser un factor de riesgo. En cambio, pertenecer al grupo de edad mayor o tener la mutación del gen TP53 siguen teniendo entre 3 y 4 veces más de riesgo en el tiempo de supervivencia, que pacientes menores o que no tengan esta mutación.

Tener una AML secundaria o presentar los genes ASXL1 o RUNX1 aproximadamente tienen 1,3 veces más de riesgo frente a los pacientes que tiene la enfermedad primaria o no poseen estas mutaciones.

Los genes BI-CEBPA y NPM1 siguen siendo favorables como se estudió previamente, ya que poseer estas mutaciones reduce en más de la mitad el riesgo en la supervivencia de los pacientes.

4. CONCLUSIONES

Para concluir el estudio se valorarán los resultados obtenidos en la estadística descriptiva de los datos, en los contrastes de hipótesis y en los análisis de supervivencia realizados a los datos de pacientes con Leucemia Mieloide Aguda, para la toma de conclusiones.

Respecto a las dos categorías clasificatorias de la enfermedad, AML primaria y **AML secundaria**, se ha confirmado que esta última tienen un peor pronóstico y una menor supervivencia respecto a los pacientes que padecen la primaria.

Se ha comprobado que la **edad** tiene una gran relevancia en la supervivencia, considerándose un factor de riesgo en pacientes mayores de 65 años, ya que aumenta hasta en 3 veces el riesgo de muerte respecto a un paciente menor a esta edad.

Al estudiar conjuntamente las categorías de AML y la edad de los pacientes, se ha observado que los pacientes mayores evolucionan peor, independientemente del tipo de AML. Tiene mejor supervivencia un paciente más joven con una AML secundaria, que uno de edad mayor con una AML primaria.

Se ha confirmado la clasificación del riesgo en favorable, intermedio y desfavorable de los pacientes por la **European LeukemiaNet 2017** a través de las anomalías genéticas que padecen los pacientes, obteniendo las peores supervivencias en los pacientes que poseen las anomalías del grupo desfavorable y las mejores en los que poseen las del favorable.

La mutación del gen **TP53** es la que mayor riesgo aporta en la supervivencia, reduciéndola en aproximadamente 3 veces respecto a un paciente que no tiene dicha mutación. Los genes **ASXL1** y **RUNX1** también reducen la supervivencia, aumentando en torno 1,3 veces el riesgo de los pacientes. Estos tres genes están considerados dentro de las anomalías citogenéticas del grupo de riesgo desfavorable según la ELN, por lo que se ha validado en nuestra base de datos dichas anomalías para la estratificación del riesgo.

Se ha comprobado de igual modo, que la mutación de los genes **BI-CEBPA** y **NPM1** hace que los pacientes evolucionen de forma favorable y aumente su supervivencia. Coincidiendo estos resultados nuevamente con que estos genes pertenecen a las anomalías del grupo de riesgo favorable según la ELN.

En los resultados obtenidos cuando los pacientes presentan conjuntamente la mutación de un gen favorable, como NPM1, y uno de riesgo, como TP53, en el trascurso de la enfermedad, prima el efecto negativo de este último, anulando toda parte positiva del primero.

Al seguir por esta línea de investigación, se ha visto que el gen **NPM1** pierde su efecto favorable y no es tan bueno en el curso de la enfermedad cuando aparece acompañado de otro gen no clasificado en los grupos de riesgo de la ELN, ni que supone un efecto negativo en los pacientes si aparece mutado exclusivamente, como **DNMT3A**.

Respecto al **género** de los pacientes, se ha visto que los hombres evolucionan peor que las mujeres con el tiempo, pero no se ha considerado como factor de riesgo ya que en los análisis realizados no ha salido como variable significativa.

En resumen, tras realizar diferentes análisis se ha visto que los factores más significativos en la supervivencia de los pacientes con Leucemia Mieloide Aguda son la edad, el tipo de AML según su orden de aparición y poseer las mutaciones de los genes NPM1, TP53, ASXL1 y RUNX1. Teniendo mayor riesgo los pacientes con TP53 mutado y con más edad.

5. BIBLIOGRAFÍA

Awada, H., Durmaz, A., Gurnari, C., Kishtagari, A., Meggendorfer, M., Kerr, C. M., Kuzmanovic, T., Durrani, J., Shreve, J., Nagata, Y., Radivoyevitch, T., Advani, A. S., Ravandi, F., Carraway, H. E., Nazha, A., Haferlach, C., Sauntharajah, Y., Scott, J., Visconte, V., … Maciejewski, J. P. (2021). Machine learning integrates genomic signatures for subclassification beyond primary and secondary acute myeloid leukemia. *Blood*, *138*(19), 1885-1895. https://doi.org/10.1182/blood.2020010603

Chacón Beltrán, R. (2008). *El uso de expresiones regulares en la detección de errores escritos: Implicaciones para el diseño de un corrector gramatical.*

De Kouchkovsky, I., & Abdul-Hay, M. (2016). 'Acute myeloid leukemia: A comprehensive review and 2016 update'. *Blood Cancer Journal*, *6*(7), e441-e441. https://doi.org/10.1038/bcj.2016.50

Devine, S. M., & Larson, R. A. (1994). Acute leukemia in adults: Recent developments in diagnosis and treatment. *CA: A Cancer Journal for Clinicians*, *44*(6), 326-352. https://doi.org/10.3322/canjclin.44.6.326

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F. R., Büchner, T., Dombret, H., Ebert, B. L., Fenaux, P., Larson, R. A., Levine, R. L., Lo-Coco, F., Naoe, T., Niederwieser, D., Ossenkoppele, G. J., Sanz, M., Sierra, J., Tallman, M. S., Tien, H.-F., … Bloomfield, C. D. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*, *129*(4), 424-447. https://doi.org/10.1182/blood-2016-08-733196

El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University - Computer and Information Sciences*, *23*(2), 91-104. https://doi.org/10.1016/j.jksuci.2011.05.005

Fisher, R. A. (1925). *043: Applications of" Student's" Distribution.*

Grimwade, D., Ivey, A., & Huntly, B. J. P. (2016). Molecular landscape of acute myeloid leukemia in younger adults and its clinical relevance. *Blood*, *127*(1), 29-41. https://doi.org/10.1182/blood-2015-07-604496

Haferlach, T., & Schmidts, I. (2020). The power and potential of integrated diagnostics in acute myeloid leukaemia. *British Journal of Haematology*, *188*(1), 36-48. https://doi.org/10.1111/bjh.16360

Hettmansperger, T., & McKean, J. (1998). Robust Nonparametric Statistical Methods. *All Books and Monographs by WMU Authors*. https://scholarworks.wmich.edu/books/579

Jayalath, K. P., Ng, H. K. T., Manage, A. B., & Riggs, K. E. (2017). Improved tests for homogeneity of variances. *Communications in Statistics - Simulation and Computation*, *46*(9), 7423-7446. https://doi.org/10.1080/03610918.2016.1241404

Kim, T. K. (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology*, *68*(6), 540-546. https://doi.org/10.4097/kjae.2015.68.6.540

Livingston, E. H. (2004). Who was student and why do we care so much about his t-test?1. *Journal of Surgical Research*, *118*(1), 58-65. https://doi.org/10.1016/j.jss.2004.02.003

McGowan-Jordan, J., Simons, A., & Schmid, M. (Eds.). (2016). *ISCN 2016: An International System for Human Cytogenomic Nomenclature (2016)*. S.Karger AG. https://doi.org/10.1159/isbn.978-3-318-06861-0

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 143-149. https://doi.org/10.11613/BM.2013.018

McKight, P. E., & Najab, J. (2010). Kruskal-Wallis Test. En *The Corsini Encyclopedia of Psychology* (pp. 1-1). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470479216.corpsy0491

Odoi, B., Twumasi-Ankrah, S., Samita, S., & Al-Hassan, S. (2022). The Efficiency of Bartlett's Test using Different forms of Residuals for Testing Homogeneity of Variance in Single and Factorial Experiments-A Simulation Study. *Scientific African*, *17*, e01323. https://doi.org/10.1016/j.sciaf.2022.e01323

Pelcovits, A., & Niroula, R. (2010). *Acute Myeloid Leukemia: A Review*.

Prada-Arismendy, J., Arroyave, J. C., & Röthlisberger, S. (2017). Molecular biomarkers in acute myeloid leukemia. *Blood Reviews*, *31*(1), 63-76. https://doi.org/10.1016/j.blre.2016.08.005

Shapiro, S. S., & Wilk, M. B. (1965). *An Analysis of Variance Test for Normality (Complete Samples)*.

Simons, A., Shaffer, L. G., & Hastings, R. J. (2013). Cytogenetic Nomenclature: Changes in the ISCN 2013 Compared to the 2009 Edition. *Cytogenetic and Genome Research*, *141*(1), 1-6. https://doi.org/10.1159/000353118

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1), 1-25. https://doi.org/10.2307/2331554

Tazi, Y., Arango-Ossa, J. E., Zhou, Y., Bernard, E., Thomas, I., Gilkes, A., Freeman, S., Pradat, Y., Johnson, S. J., Hills, R., Dillon, R., Levine, M. F., Leongamornlert, D., Butler, A., Ganser, A., Bullinger, L., Döhner, K., Ottmann, O., Adams, R., … Papaemmanuil, E. (2022). Unified classification and risk-stratification in Acute Myeloid Leukemia. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-32103-8